MODELING INTENTION, EMOTION AND EXTERNAL WORLD IN DIALOGUE SYSTEMS

Wei Peng [1,2], *Yue Hu*[1,2]*, *Luxi Xing*[1,2], *Yuqiang Xie*[1,2], *Xingsheng Zhang*[1,2], *Yajing Sun*[1,2]

[1]Institute of Information Engineering, Chinese Academy of Sciences, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, China

ABSTRACT

Intention, emotion and action are important elements in human activities. Modeling the interaction process between individuals by analyzing the relationships between these elements is a challenging task. However, previous work mainly focused on modeling intention and emotion independently, and neglected of exploring the mutual relationships between intention and emotion. In this paper, we propose a RelAtion Interaction Network (RAIN), consisting of Intention Relation Module and Emotion Relation Module, to jointly model mutual relationships and explicitly integrate historical intention information. The experiments on the dataset show that our model can take full advantage of the intention, emotion and action between individuals and achieve a remarkable improvement over BERT-style baselines. Qualitative analysis verifies the importance of the mutual interaction between the intention and emotion.

Index Terms— Human Interaction, Intention Recognition, Emotion Prediction

1. INTRODUCTION

Intention recognition and emotion prediction are long-term researches in dialogue systems [1, 2]. Intention [3, 4], as an essential psychological background to stimulate and guide action, is the internal dynamic tendency to carry out activities, thus affecting the state of the external world. For example, when a person wants to purchase, he will have a conversation with the shop assistant to ask to know the information about the goods, including material, price, size, etc. At this time, the intention of the customer can be also inferred to purchase from the act of dialogue. On the contrary, he will not communicate with the shop assistant if the customer has no intention.

When the state of the external world changes (the customer bought satisfactory goods) and the change satisfies the intention, it is intuitive the customer is happy. As the work [5] demonstrated, emotion is the psychological behavior produced by the joint stimulation of the internal and external world. Namely, emotion is determined by both their own intention and external action. It often shows positive emotion when the action satisfies the intention [5, 6]. Figure 1 shows the relationships between the intention, emotion and the action between the speaker and listener. Person A is thirst and wants to drink, so he puts forward a request to satisfy his intention *request for drinking*. From the current intention and utterance, his emotion can be inferred as awful. Then, a new action *Do you want orange or Coca-Cola* is triggered by person B from which the intention is inferred to *question*. Similarly, person A expresses a new intention *inform* by the utterance *Coca-Cola, please*. And the emotion has been changed to *happy* because of the satisfaction of historical intention *request for drinking*. Thus, it's critical to take the mutual relationships between the intention and emotion into account in an explicit way.

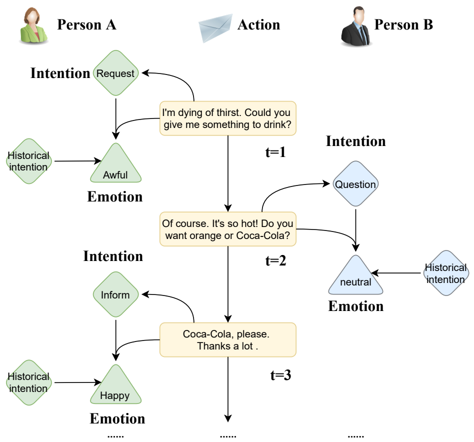


Fig. 1. An example of dialogue systems. Diamond squares indicate intention, and different colors indicate different individuals. The triangle represents the emotion, and the rectangle represents the interaction action in the external world.

Previous work [7, 8, 9] mainly focused on intention recognition or emotion prediction, but seldom considered these factors together and ignored to explicitly integrate historical intention information. Although some work [10, 11, 12] proposed modeling the interaction between the intention recognition task and emotion prediction task, they just predicted the label by the representation of utterances or self attention mechanism and regarded the two tasks as the sentence classification task. These works also lack analysis and explanation of intention recognition and emotion prediction.

*Corresponding author. This work is supported by the National Natural Science Foundation of China (U21B2009).

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

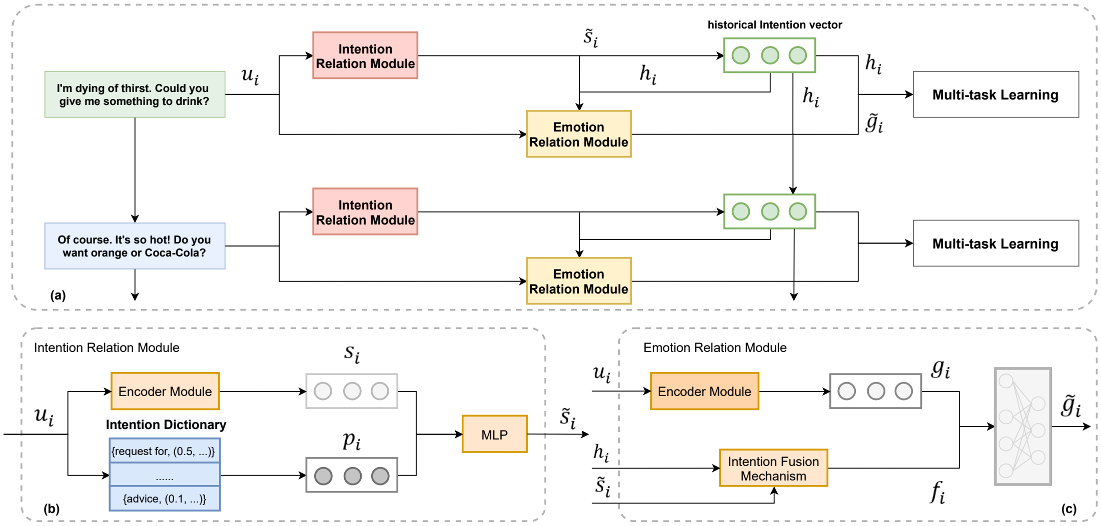


Fig. 2. (a) is the overview of our RAIN which consists of the Intention Relation Module (b) and Emotion Relation Module (c).

In this paper, we propose a **RelAtion Interaction Network (RAIN)**, consisting of Intention Relation Module and Emotion Relation Module, to jointly model mutual relationships and explicitly integrate historical intention information. Specifically, **Intention Relation Module** introduces an intention dictionary to explicitly account for the intention recognition task. Then, **Emotion Relation Module** designs an intention fusion mechanism to explicitly integrate historical intention information for subsequent emotion prediction. The important observation is the significant performance obtained from RAIN; it not only achieves the high performance on the dataset but also makes an explanation of the two tasks.

2. METHODOLOGY

As shown in Figure 2 (a), the proposed model consists of the Intention Relation Module and Emotion Relation Module. First, the **Intention Relation Module** (b) obtains the contextual representations of dialogues with the econder module and introduces an intention dictionary to make an interpretable intention recognition. Then, the **Emotion Relation Module** (c) designs an intention fusion mechanism to explicitly integrate historical intention information. Finally, RAIN makes subsequent predictions with the intention vectors $\tilde{\boldsymbol{s}}_{\boldsymbol{i}}$ and emotion vectors $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}$ by multi-task learning.

2.1. Intention Relation Module

Encoder Module. Considering the strong performance of pre-trained language models (PLMs) [13, 14, 15] such as RoBERTa, we use them as the Encoder Module to obtain the contextual representations. Given a conversation $C = (u_1, u_2, \ldots, u_N)$ a set of N utterances and $u_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,T})$ that consists of a sequence of T words, with $Y^s = (y_1^s, y_2^s, \ldots, y_N^s)$ and $Y^e = (y_1^e, y_2^e, \ldots, y_N^e)$ being the corresponding intention and emotion labels. To obtain the i^{th} sentence representation, we consider using the hidden state corresponding to [CLS], as:

$$\boldsymbol{s}_{\boldsymbol{i}} = \text{RoBERTa}\left([CLS], x_{i,1}, \ldots, x_{i,T}, [SEP]\right)[0] \quad (1)$$

Intention Dictionary. Observing that certain specific words such as *ask for, proposal* can reflect intention, we extract the feature words in the process of labeling the feature. For example, *request for, would like* has a high probability of being *request*. We count the number of labeled feature words, count the word frequency of each feature word over different intentions, and normalize it to get the probability distribution $\boldsymbol{p}_{\boldsymbol{i}}$. Therefore, we construct an Intention Dictionary as the prior intention knowledge base (Figure 2 (b)) and then output a probability distribution $\boldsymbol{p}_{\boldsymbol{i}}$ of the keyword over all the corresponding intentions, which can be regarded as an interpretable signal to enhance the semantic of the intention.

Finally, the intention vectors $\tilde{\boldsymbol{s}}_{\boldsymbol{i}}$, integrating symbolic representation and neural representation, can be obtained as:

$$\tilde{\boldsymbol{s}}_{\boldsymbol{i}} = \text{MLP}\left(\text{ReLU}(W_s^T \boldsymbol{s}_{\boldsymbol{i}} + \boldsymbol{p}_{\boldsymbol{i}})\right) \quad (2)$$

where $\tilde{\boldsymbol{s}}_{\boldsymbol{i}} \in \mathbb{R}^h$, $W_s \in \mathbb{R}^{h \times l_s}$, h is the dimension of the hidden state, l_s is the number of the intention labels.

2.2. Emotion Relation Module

In this module, the inputs includes: the current utterance u_i, the historical intention information $\boldsymbol{h}_{\boldsymbol{i}}$ and the output of the Intention Relation Module $\tilde{\boldsymbol{s}}_{\boldsymbol{i}}$. The details are as follows.

Encoder Module. As shown in Figure 2 (c), the Emotion Relation Module utilizes another RoBERTa [15] as the Encoder Module for obtaining the sentence-level representation $\boldsymbol{g}_{\boldsymbol{i}}$ with the same formulation in Equation (1), where $\boldsymbol{g}_{\boldsymbol{i}} \in \mathbb{R}^h$.

	Intention Recognition			Emotion Prediction		
	P ↑	R ↑	F1 ↑	P ↑	R ↑	F1 ↑
GRU [17]	46.18	41.31	43.61	42.25	41.64	41.94
GRU+Attention [13]	48.66	41.43	44.75	43.74	41.20	42.43
DCR-Net † [12]	52.35	48.56	50.38	47.24	43.91	45.51
BERT [18]	65.84	65.18	65.51	55.35	55.42	55.38
RoBERTa$_{base}$ [15]	68.14	68.53	68.33	56.82	57.89	57.35
RoBERTa$_{large}$ [15]	71.36	70.47	70.91	59.51	58.77	59.13
RAIN	**73.22**	**72.64**	**72.93**	**65.35**	**62.84**	**64.07**

Table 1. Experimental results on the testset for tasks of intention recognition and emotion prediction. † indicates that the performance is reimplemented by ourselves.

Historical Intention Modeling. As demonstrated in Section 1, emotion is the psychological behavior produced by the joint stimulation of the intention and action. To model mutual relationships between the intention and emotion, the proposed model introduces historical intention information $\boldsymbol{h_i}$. Considering the sequence modeling, we utilize the LSTM [16] to capture the temporal features within the intentions, as:

$$\boldsymbol{h_i} = \text{LSTM}\ (\tilde{\boldsymbol{s}}_{\boldsymbol{i}}, \boldsymbol{h_{i-1}})) \tag{3}$$

Intention Fusion Mechanism. The fusion mechanism [19, 20] is a general approach that is model-independent, which focus on the mutual relationships between the two sources. Motivated by the work [21], the intention fusion mechanism is proposed to effectively integrate historical intention information. Specifically, the fusion layer first utilizes a particular fusion unit to combine the representations between the intention vectors $\tilde{\boldsymbol{s}}_{\boldsymbol{i}}$ and the historical intention information $\boldsymbol{h_i}$.

$$\boldsymbol{f_i} = \text{Fuse}\ (\tilde{\boldsymbol{s}}_{\boldsymbol{i}}, \boldsymbol{h_i}) \tag{4}$$

where $\text{Fuse}(\cdot, \cdot)$ is a typical fusion kernel.

The simplest fusion kernel can be a concatenation or addition operation of the two sources, followed by a linear or non-linear transformation. Generally, a heuristic matching trick with difference and element-wise product is found effective in combining different representations [19, 20]:

$$\text{Fuse}(\tilde{\boldsymbol{s}}_{\boldsymbol{i}}, \boldsymbol{h_i}) = \tanh(W_f^T[\tilde{\boldsymbol{s}}_{\boldsymbol{i}}; \boldsymbol{h_i}; \tilde{\boldsymbol{s}}_{\boldsymbol{i}} \circ \boldsymbol{h_i}; \tilde{\boldsymbol{s}}_{\boldsymbol{i}} - \boldsymbol{h_i}] + b_f) \tag{5}$$

where $W_f \in \mathbb{R}^{4h \times h}$, $b_f \in \mathbb{R}^h$ are trainable parameters, $\circ$ denotes the element-wise product. The output dimension is projected back to the same size as the original representation $\tilde{\boldsymbol{s}}_{\boldsymbol{i}}$ or $\boldsymbol{h_i}$. [;] indicates vector concatenation.

Finally, the emotion vectors $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}$, modeling the mutual relationships between the intention and emotion, obtained as:

$$\tilde{\boldsymbol{g}}_{\boldsymbol{i}} = \text{ReLU}\ (W_g^T[\boldsymbol{f_i}; \boldsymbol{g_i}] + b_g) \tag{6}$$

where $\tilde{\boldsymbol{g}}_{\boldsymbol{i}} \in \mathbb{R}^h$, $W_g \in \mathbb{R}^{2h \times h}$ and $b_g \in \mathbb{R}^h$.

2.3. Prediction for Intention and Emotion

Intention Relation Module integrates symbolic representation and neural representation to output intention vectors $\tilde{\boldsymbol{s}}_{\boldsymbol{i}}$. Emotion Relation Module models the mutual relationships between the intention and emotion, and fuses the historical intention information to output emotion vectors $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}$. After obtaining the $\tilde{\boldsymbol{s}}_{\boldsymbol{i}}$ and $\tilde{\boldsymbol{g}}_{\boldsymbol{i}}$, we then consider two MLPs for performing intention recognition and emotion prediction, which is defined as:

	Intention Recognition		Emotion Prediction	
	F1 ↑	$\Delta_{(F1)}$	F1 ↑	$\Delta_{(F1)}$
RoBERTa$_{large}$ [15]	70.91	-	59.13	-
+IntentNet	72.29	+1.38	61.15	+2.02
+Fusion Mechanism	-	-	60.46	+1.33
+Historical Intention	71.77	+0.86	62.11	+2.98
+Multi-task	71.96	+1.05	59.97	+0.84
RAIN	**72.93**	**+2.02**	**64.07**	**+4.94**

Table 2. The results of ablation study on model components.

$$\hat{\boldsymbol{y}}_i^m = \text{Softmax}(W_m^T \tilde{\boldsymbol{s}}_{\boldsymbol{i}} + b_m), \tag{7}$$

$$\hat{\boldsymbol{y}}_i^e = \text{Softmax}(W_e^T \tilde{\boldsymbol{g}}_{\boldsymbol{i}} + b_e), \tag{8}$$

where $\hat{\boldsymbol{y}}_i^m$, $\hat{\boldsymbol{y}}_i^e$ are the predicted distribution for intention and emotion, $W_m^T \in \mathbb{R}^{h \times l_m}$, $W_e^T \in \mathbb{R}^{h \times l_e}$ are transformation matrices, l_m, l_e are the number of intention and emotion labels.

The average cross-entropy loss of the intention recognition and emotion prediction are optimized as:

$$\mathcal{L}_m = -\frac{1}{K}\sum_{j=1}^{K}\sum_{i=1}^{N} y_i^m \log \hat{\boldsymbol{y}}_i^m \tag{9}$$

$$\mathcal{L}_e = -\frac{1}{K}\sum_{j=1}^{K}\sum_{i=1}^{N} y_i^e \log \hat{\boldsymbol{y}}_i^e \tag{10}$$

where K is the total number of the examples, N is the number of the utterances in one conversation, y_i^m and y_i^e are gold utterance intention label and gold emotion label.

Joint Learning. We combine the above two loss functions as the training loss in a multi-task learning manner as:

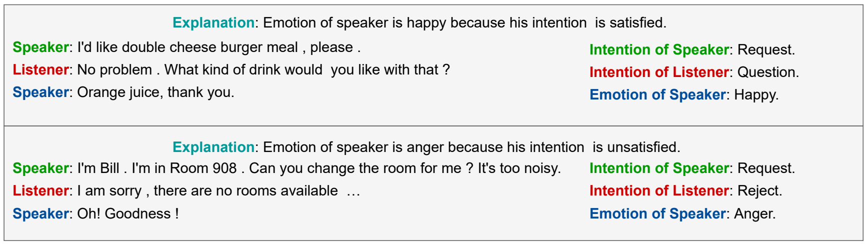
$$\mathcal{L}(\theta) = \lambda_1 \mathcal{L}_m + \lambda_2 \mathcal{L}_e \tag{11}$$

where θ is the all learnable parameters, and λ_1 and λ_2 are two hyper-parameters for controlling the weight of the rest tasks.

3. EXPERIMENTS

3.1. Experimental Setting

Datasets & Evaluation Metrics. The DailyDialog [22] datasets contain 13,118 multi-turn dialogues. In order to obtain the intention dictionary, we extract 2,046 conversations for the annotation. Furthermore, the intention classification has been expanded into seven categories (four categories in the original dataset), including *request, suggest, command, accept, reject, question* and *inform*. The emotion classification including *happy, neutral, sadness, anger, content* and *disgust*. As for the evaluation metric, macro-average Precision (P), Recall (R) and F1 are considered for the DailyDialog dataset.



Explanation: Emotion of speaker is happy because his intention is satisfied.

Speaker: I'd like double cheese burger meal , please .
Listener: No problem . What kind of drink would you like with that ?
Speaker: Orange juice, thank you.

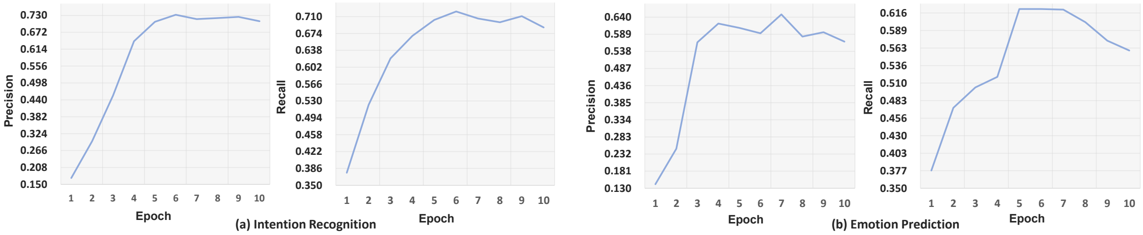
Intention of Speaker: Request.
Intention of Listener: Question.
Emotion of Speaker: Happy.

Explanation: Emotion of speaker is anger because his intention is unsatisfied.

Speaker: I'm Bill . I'm in Room 908 . Can you change the room for me ? It's too noisy.
Listener: I am sorry , there are no rooms available …
Speaker: Oh! Goodness !

Intention of Speaker: Request.
Intention of Listener: Reject.
Emotion of Speaker: Anger.

Fig. 3. Two examples RAIN generated, we also give the explanation of the emotion prediction task.



(a) Intention Recognition

(b) Emotion Prediction

Fig. 4. The performance with respect to the number of epochs on the dataset.

Implementation details. The BERT-style baselines have same hyper parameters as [15]. Additionly, we perform a grid search over the hyper-parameter settings (with a learning rate from $\{0.01, 0.4\}$ for GRU or $\{$1e-5, 3e-5$\}$ for PLMs (in Sec. 2.1), a batch size from $\{16, 32\}$, and epochs from $\{3, 12\}$). The hyper-parameters in the loss are $\lambda_1 = \lambda_2 = 0.5$. Models are trained to minimize the cross entropy with Adam [23].

3.2. Experimental Results

State-of-the-art Comparison. We provide some baselines and proposed model for the two tasks. For the intention recognition, as shown in the first column of Table 1, in all models, RoBERTa performs the best. Compared with the baselines, RAIN can outperform them by a large margin which achieves 2.02% gain on F1. For the emotion prediction, as shown in the second column of Table 1, RAIN achieves the best results that outperform the previous best-published model.

Ablation Study. To get a better insight into our proposed model, we perform the ablation study as shown in Table 2. We add the proposed modules to explore their contribution. For comparison, the RoBERTa$_{large}$ baseline and ours that utilizes all the components are also provided.

From the Table 2, we can see that: 1) After utilizing the intention dictionary, the performance has improved on the two tasks. It shows that the prior statistical knowledge is effective to predict the intention as well as the emotion. 2) The fusion mechanism also has an improvement, proving that such an operation is indeed effective. 3) The historical intention information is beneficial for the two tasks, which validate the necessity of the historical modeling. 4) The multi-task learning can provide benefits, which show that the two training objectives are actually closely related and can boost each other. The ablation study has demonstrated that all the components proposed are beneficial for the tasks.

Case Study. In this section, we present dialogue cases and explanations of the emotion to demonstrate how our model performs. As shown in Fig. 3, for the first case, the speaker requests for *burger meal* and intention of the listener is *question*. Our model produces a reasonable intention and emotion prediction. As for the explanation of the emotion, we give the generative template based on the outputs, like *Emotion of speaker is happy because his intention is satisfied*.

Performance on the Number of Epochs. To explore the influence of different epochs, another experiment with the proposed model is conducted. The comparison of the results is depicted in Fig. 4. From the experiments, the conclusion is that during the first $\{5, 6\}$ epochs, performances achieve a lot and reach a peak, while after that, performances drop slightly. It demonstrates that the models can converge quickly during the first few epochs and capture the features between the utterances effectively.

4. CONCLUSION

In this paper, we present a RelAtion Interaction Network (RAIN) to jointly model mutual relationships between the intention and emotion, and explicitly integrate historical intention information. We show that the proposed RAIN is effective and interpretable, which outperforms the previous methods with a single model, as well as making an explanation of the two tasks. For the future work, some other psychological factors will be considered, such as personal character, educational background and so on. We believe that these cognitive factors are still worth researching for human activities.

5. REFERENCES

[1] Rosalind W. Picard, *Affective computing*, MIT Press, 1997.

[2] Abraham Harold Maslow, *A theory of human motivation*, Simon and Schuster, 2013.

[3] Philip A. Gable and Eddie Harmon-Jones, "Approach-motivated positive affect reduces breadth of attention," *Psychological Science*, vol. 19, no. 5, pp. 476–482, 2010.

[4] Johnmarshall Reeve, *Understanding motivation and emotion*, John Wiley & Sons, 2014.

[5] Joseph J. Campos, Rosemary G. Campos, and Karen C. Barrett, "Emergent themes in the study of emotional development and emotion regulation.," *Developmental Psychology*, vol. 25, no. 3, pp. 394–402, 1989.

[6] E. J. Lawrence, P. Shaw, V. P. Giampietro, S. Surguladze, M. J. Brammer, and A. S. David, "The role of 'shared representations' in social perception and empathy: an fmri study.," *Neuroimage*, vol. 29, no. 4, pp. 1173–1184, 2006.

[7] Harshit Kumar, Arvind Agarwal, and Sachindra Joshi, "A practical dialogue-act-driven conversation model for multi-turn response selection," in *EMNLP-IJCNLP*, 2019, pp. 1980–1989.

[8] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria, "Dialoguernn: An attentive RNN for emotion detection in conversations," in *AAAI*, 2019, pp. 6818–6825.

[9] Pierre Colombo, Emile Chapuis, Matteo Manica, Emmanuel Vignon, Giovanna Varni, and Chloé Clavel, "Guiding attention in sequence-to-sequence models for dialogue act prediction," in *AAAI*, 2020, pp. 7594–7601.

[10] Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le, "Multi-task dialog act and sentiment recognition on mastodon," in *COLING*, 2018, pp. 745–754.

[11] Minkyoung Kim and Harksoo Kim, "Integrated neural network model for identifying speech acts, predicators, and sentiments of dialogue utterances," *Pattern Recognit. Lett.*, vol. 101, pp. 1–5, 2018.

[12] Libo Qin, Wanxiang Che, Yangming Li, Minheng Ni, and Ting Liu, "Dcr-net: A deep co-interactive relation network for joint dialog act recognition and sentiment classification," in *AAAI*, 2020, pp. 8665–8672.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, L. Kaiser, and Illia Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[14] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," *CoRR*, vol. abs/1909.11942, 2019.

[15] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, pp. 1735–1780, 1997.

[17] Junyoung Chung, Çaglar Gülçehre, KyungHyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *CoRR*, vol. abs/1412.3555, 2014.

[18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT*, 2019, pp. 4171–4186.

[19] Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Jing Yu, Yajing Sun, and Xiangpeng Wei, "Bi-directional cognitive-thinking network for machine reading comprehension," in *COLING*, Donia Scott, Núria Bel, and Chengqing Zong, Eds. 2020, pp. 2613–2623, International Committee on Computational Linguistics.

[20] Lili Mou, Rui Men, Ge Li, Yan Xu, Lu Zhang, Rui Yan, and Zhi Jin, "Natural language inference by tree-based convolution and heuristic matching," in *ACL*. 2016, The Association for Computer Linguistics.

[21] Wei Peng, Yue Hu, Jing Yu, Luxi Xing, Yuqiang Xie, Zihao Zhu, and Yajing Sun, "MCR-NET: A multi-step co-interactive relation network for unanswerable questions on machine reading comprehension," in *ICASS*. 2021, pp. 7818–7822, IEEE.

[22] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *IJCNLP*, 2017, pp. 986–995.

[23] Diederik P. Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.