ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746537

ADJACENCY PAIRS-AWARE HIERARCHICAL ATTENTION NETWORKS FOR DIALOGUE INTENT CLASSIFICATION

Jiabao Xu, Peijie Huang, Youming Peng, Jiande Ding, Boxi Huang, Simin Huang*

College of Mathematics and Informatics, South China Agricultural University, China
jiabao0668@foxmail.com, pjhuang@scau.edu.cn,ympeng@stu.scau.edu.cn,
jiandeding1999@163.com,{bokheiwong, huangsimin}@stu.scau.edu.cn

ABSTRACT

Dialogue intent classification is a fundamental and essential task in dialogue systems. Although sentence-level and document-level text classification have made dramatic progress in recent years with the help of deep learning technology, dialogue-level classification remains challenging. Dialogue has unique characteristics that distinguish it from other types of text. Dialogue is interactive, with feedback between speakers, and turn-taking. These unique features suggest that model architecture should take dialogue structure into account to learn a better representation. In this paper we propose an Adjacency Pairs-Aware Hierarchical Attention Network (AP-HAN) for dialogue intent classification. A dialogue reconstruction strategy is designed to match the question and answer utterances properly and then make the dialogue to be presented as a sequence of adjacent pairs. Then, the adjacency pairs features are incorporated into the hierarchical attention network. Experimental results on public CCL2018-Task1 corpus show the better performance of the proposed model.

Index Terms— Intent classification, Dialogue modeling, Adjacency pairs, Hierarchical attention network.

1. INTRODUCTION

Intent classification is a fundamental and essential task in dialogue systems. Using automatic speech recognition (ASR) to transcribe user speech input into text, and then classify it into a predefined label is the usual method to achieve the purpose of recognizing and understanding user intent of a dialogue.

Dialogue intent classification is one kind of text classification. Traditionally, shallow models, such as SVM [1] and logistics regression, are used to learn text representation for classification. In recent years, deep learning models, such as CNN [2], RNN [3], attention mechanism [4] and their hybrids are widely used. Sentence-level and document-level text classification have made dramatic progress in recent years [5–7].

Speaker	Turns	Dialogue fragment
A	T1	Hello very glad to serve you!
B	T2	Hello, why can't I use the ten yuan domestic package I just ordered?
A	T3	It has taken effect. You can use it directly online now.
B	T4	Can I use it directly online now.
A	T5	Yes, the order has been successful at 10:14.

Table 1. An example of dialogue fragment from the CCL2018-Task1 corpus

Nonetheless, dialogue-level classification remains challenging and relatively under-investigated. Dialogue has unique characteristics that distinguish it from other types of text. Dialogue is interactive in nature, with feedback between speakers, and turn-taking [8]. Further, important pieces of information may be scattered across various utterances of different speakers and capturing the intent behind them requires deeper understanding of the dialogue context. These unique features suggest that model architecture should take dialogue structure into account to learn a better representation.

Several studies have been proposed to integrate dialogue structure to learn dialogue context representation on pre-trained language model [9], Multi-turn Response Selection [10], and dialogue summarization [11]. Although these neural network based approaches have been quite effective, they have not fully exploited the dialogue structure. Conversation is a cooperative language communication activity involving two persons, and information contained in a single utterance is usually incomplete. Using adjacent pairs as the basic unit of dialogue is conducive to learning more meaningful representations. Adjacent pair is made up of two talkers each speaking once [12]. As shown in Table 1, T1 is a turn and it can form an adjacent pair with T2. However, the adjacent pair isn't naturely fixed in the order of appearance of the two utterances. For example, the five utterances in Table 1 can form three adjacent pairs of T1-T2, T2-T3, and T4-T5. The wrong constructed adjacent pair may introduce additional noise, resulting in a model performance degradation.

To deal with the above challenge, we propose an Adjacency Pairs-Aware Hierarchical Attention Network (AP-HAN) for the task of dialogue intent classification. Com-

* Corresponding author.

pared to other text understanding models, our AP-HAN can better capture the unique character of dialogue. First, to model the relationships between questions and answers, we extend the work of hierarchical attention network (HAN) [7] and incorporate the structure of adjacency pairs into the neural network. Furthermore, in order to make the dialogue to be presented as a sequence of adjacent pairs, we propose a dialogue reconstruction strategy, which helps to match the questions and answers properly and reduces the negative impact of the wrong adjacent pairs. Experimental results on the public CCL2018-Task1 corpus show that the proposed model achieves significant improvements over the compared models, especially for long dialogues.

2. MODEL

This section describes our AP-HAN model (Show in Figure 1) that construct and incorporates adjacency pairs features into the hierarchical attention network.

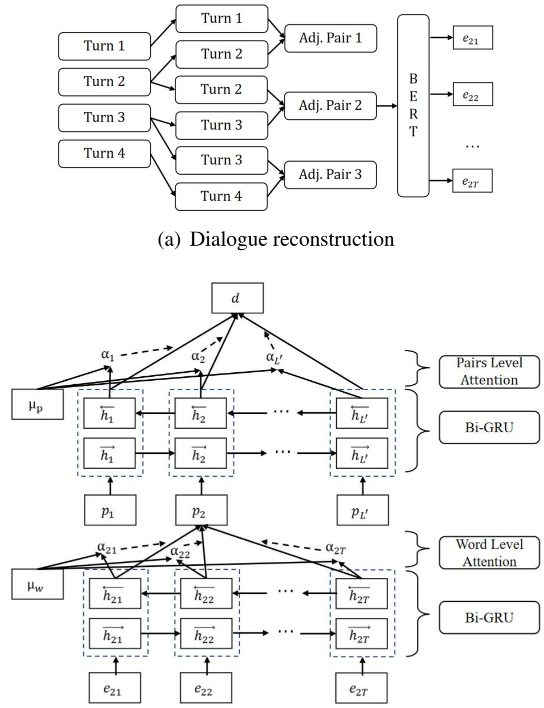
2.1. Dialogue Reconstruction

According to Schegloff and Sacks's definition [12], adjacency pair is a pair made up of two speakers each speaking once. The pair of question-answer is the most common type, especially in the field of customer service. Unfortunately, if adjacent pairs are formed according to each two turns in chronological order, it will lead to wrong adjacent pairs, which will have a negative impact on the model. Therefore, we propose dialogue reconstruction strategy to use interrogative sentence recognition and replication to make multi-turn dialogue into adjacent pairs. Formally, given a dialogue D with L utterances, our goal is to construct $L^{'}$ adjacent pairs. From the beginning of the original dialogue, every two turns will form an adjacency pair in chronological order. When the second turn of the adjacent pair is an interrogative sentence (e.g. T2 in Table 1), it will be copied once, and the copied utterance will form a new adjacent pair with the next utterance (e.g. T2-T3 in Table 1). In this paper, we follow [13] to collect interrogative words and use regular expressions to achieve interrogative sentence recognition.

Through the above strategy, we obtain the adjacent pairs with two utterance as a group, $D = \{p_1, p_2, ..., p_{L'}\}$, where $p_i = \{w_{i1}, w_{i2}, ..., w_{iT}\}$, w_{it} with $t \in [1, T]$ represents the word t in the i^{th} adjacency pair. Then we use BERT [14] to embed the words to vectors through the last layer of it and the BERT model is fine-tuned with our framework. Specifically, an adjacent pair with two turns A, B is constructed as the following input form:

$$[CLS]A[SEP]B[SEP] \tag{1}$$

This is consistent with the next sentence prediction (NSP) pre-training task of BERT model, which aims to let the model learn the correlation between utterances.



(a) Dialogue reconstruction

(b) Adjacency pairs-aware hierarchical attention network

Fig. 1. The architecture AP-HAN

2.2. Adjacency Pairs-Aware HAN

Word Encoder. Given the adjacency pair p_i with word embedding $\boldsymbol{e}_{it}, t \in [1, T]$, we employ a Bi-directional GRU [15], which can efficiently make use of past features and future features for a specific time step, to get representations of words by summarizing information from both directions. The forward GRU reads p_i from the $\boldsymbol{e}_{i1}$ to $\boldsymbol{e}_{iT}$ and a backward GRU reads from $\boldsymbol{e}_{iT}$ to $\boldsymbol{e}_{i1}$.

$$\overrightarrow{GRU}(\boldsymbol{e}_{i1}, \boldsymbol{e}_{i2}, ..., \boldsymbol{e}_{iT}) = (\overrightarrow{\boldsymbol{h}}_{i1}, \overrightarrow{\boldsymbol{h}}_{i2}, ..., \overrightarrow{\boldsymbol{h}}_{iT}) \tag{2}$$

$$\overleftarrow{GRU}(\boldsymbol{e}_{iT}, ..., \boldsymbol{e}_{i2}, \boldsymbol{e}_{i1}) = (\overleftarrow{\boldsymbol{h}}_{iT}, ..., \overleftarrow{\boldsymbol{h}}_{i2}, \overleftarrow{\boldsymbol{h}}_{i1}) \tag{3}$$

We obtain a contextual representation for a given word $\boldsymbol{e}_{it}$ by concatenating the forward and backward hidden states:

$$\boldsymbol{h}_{it} = [\overrightarrow{\boldsymbol{h}}_{it}, \overleftarrow{\boldsymbol{h}}_{it}] \tag{4}$$

which summarizes the information of the whole adjacency pair centered around the word $\boldsymbol{e}_{it}$.

Word Attention Layer. Word attention mechanism is introduced to capture which words that are important to the meaning of the pair and aggregate the representation of those informative words to form a pair vector $\boldsymbol{z}_i$:

$$\boldsymbol{u}_{it} = \tanh(\boldsymbol{W}_w \boldsymbol{h}_{it} + \boldsymbol{b}_w) \tag{5}$$

$$\alpha_{it} = \frac{exp(\boldsymbol{u}_{it}^{\top} \boldsymbol{u}_w)}{\sum_t exp(\boldsymbol{u}_{it}^{\top} \boldsymbol{u}_w)} \tag{6}$$

$$z_i = \sum_t \alpha_{it} h_{it} \quad (7)$$

where h_{it} denotes the hidden state of word t in adjacency pair i. α_{it} is the corresponding attention weight calculated by a softmax function and W_w, b_w, $u_w$are model parameters. Then we compute the adjacency pair vector z_i as a weighted sum of the word annotations based on the weights. After performing the same computation in all adjacent pairs, we get the vector sequence (i.e., $(z_1, z_2, ..., z_{L'})$).

Adjacency Pair Encoder. Given the pair vectors z_i, we also use a Bi-GRU to encode the pairs in order to incorporate the contextual information in the annotations, i.e.,

$$\overrightarrow{GRU}(z_1, z_2, ..., z_{L'}) = (\overrightarrow{h}_1, \overrightarrow{h}_2, ..., \overrightarrow{h}_{L'}) \quad (8)$$

$$\overleftarrow{GRU}(z_{L'}, ..., z_2, z_1) = (\overleftarrow{h}_{L'}, ..., \overleftarrow{h}_2, \overleftarrow{h}_1) \quad (9)$$

$$h_i = [\overrightarrow{h}_i, \overleftarrow{h}_i] \quad (10)$$

We concatenate $\overrightarrow{h}_i$ and $\overleftarrow{h}_i$ to get an annotation of pair p_i, h_i summarizes the neighbor information near pair p_i.

Adjacency Pair Attention Layer. Following [7], to reward adjacency pairs that are clues to classify a dialogue, we again use an attention mechanism and introduce a pair level context vector u_p to measure the importance of all pairs.

$$u_i = \tanh(W_p h_i + b_p) \quad (11)$$

$$\alpha_i = \frac{exp(u_i^\top u_p)}{\sum_t exp(u_i^\top u_p)} \quad (12)$$

$$d = \sum_i \alpha_i h_i \quad (13)$$

where d is the dialogue vector that summarizes all the information of adjacency pairs. Similarly, the context vector u_p can be randomly initialized and jointly learned during the training process.

2.3. Dialogue Intent Classification

Finally, we feed the dialogue representation vector into a multi-layer perceptron (MLP) and send the output of MLP to a softmax function to predict the probability of each category. To alleviate the overfitting problem, we apply dropout regularization [16]. We use cross-entropy loss function to train our model end-to-end given a set of training data $\{D_i, y_i\}, i \in [1, N]$, where D_i is the i^{th} dialogue to be predicted and y_i is the ground-truth intent category for dialogue D_i. The goal of training is to minimize the loss function:

$$\text{prob} = \text{softmax}(W_c * (r \odot d) + b_c) \quad (14)$$

$$\text{loss} = -\sum_{n=1}^{N}\sum_{k=1}^{K} y_n^k * \log(\text{prob}) \quad (15)$$

where N is the number of training samples and K is the category number. r is a vector with the same dimension as d and obeys the Bernoulli distribution. $\odot$ represents Hadamard product.

3. EXPERIMENTS

3.1. Dataset

Dataset 1. We test our model on a public CCL2018-Task1 corpus[1]. The dataset is in Chinese and contains 20000 samples with 34 intent labels, which are text data transcribed by the dialogue recording between users and customer service. According to the service type, the 34 intent labels belong to three large categories: consultation, complaint and handling. Following [17], 20% of the dataset is used as test set, and the remaining data is divided into training and validation sets according to the proportion of 8:2.

Dataset 2. In addition, we further divide dataset 1 into three data subsets of the individual large categories, and conduct experiments on each of them, respectively. It can eliminate the internal confusion of the three large service categories, and be used for the evaluation of intention recognition of the small categories within each large category.

3.2. Experiment Settings

The hyperparameters of the models are tuned on the validation set. We set the dimension of the hidden states of GRU as 300. To avoid overfitting, dropout [16] with a probability of 0.2 is used. For training, we use a mini-batch size of 24 and dialogues with the same numbers of adjacency pairs are organized to be a batch. The parameters are updated by the Adam algorithm [18] and the learning rate is initialized as 2e-5.

We compare our model with several baseline methods: (1) **BERT Fine-Tune**: A pre-trained language model [14] for text classification. (2) **BiLSTM**: A classic baseline that is widely used for text classification [19]. (3) **BiLSTM Soft ATT**: A standard BiLSTM with soft attention mechanism for text classification [20]. (4) **HAN**: A hierarchical network model with both word- and sentence-level attentions proposed by Yang et al. [7]. (5) **PLA-HAN**: Another hierarchical network model with both word- and sentence-level attentions proposed by Ding et al. [17]. They incorporate utterance label attention using an auxiliary external data set. For the fair experiments, all of the baselines and our AP-HAN are built on the top of **BERT**, and process fine-tuning during training.

3.3. Main Results

Table 2 shows the result of our AP-HAN and competing approaches: (1) Compared with the sequence models that take the whole dialogue text as input, hierarchical structure models achieve better classification performance. The hierarchical models can better model the semantic structure of words, sentences and dialogue. It may also because hierarchical attention can avoid the maximum sequence length limitation when using BERT. (2) By incorporating utterance attention using

[1] http://www.cips-cl.org/static/CCL2018/call-evaluation.html

Model types	Models	Dataset 1	Dataset 2		
			Consultation	Handling	Complaint
Sequence model	BERT FineTune	53.11	59.23	93.50	67.66
	BiLSTM	53.75	58.27	93.62	66.71
	BiLSTM Soft ATT	55.87	58.32	93.62	67.52
Hierarchical model	HAN	56.31	58.48	93.85	67.52
	PLA-HAN	56.94	58.89	93.92	67.41
	AP-HAN	**57.60***	**60.15***	**94.66***	**69.28***
	w/o dialogue reconstruction	56.28	59.52	94.43	68.34

Table 2. Comparison of intent accuracy (%) of our model with baselines on test datasets. The numbers with * indicate that the improvement of our model over all baselines is statistically significant with $p < 0.01$ under t-test

an auxiliary external data set, PLA-HAN outperforms HAN. But PLA-HAN also has not fully exploited the dialogue structure. Our AP-HAN, which reconstructs the dialogue to be adjacency pairs-aware, achieves the best performance, which is better than HAN and PLA-HAN models in both datasets. (3) Experiment of the effect on dialogue reconstruction (i.e. AP-HAN w/o dialogue reconstruction) verifies the effectiveness of the dialogue reconstruction strategy in reducing the negative impact of wrong adjacent pairs.

3.4. Further Analysis

Impact on Dialogue Lengths. We further compare the performance of AP-HAN with HAN and PLA-HAN under different dialogue lengths using dataset 1. We divide it into three categories: long (more than 600 Chinese characters), medium (301-600 Chinese characters) and short (less than 300 Chinese characters). The results are shown in Table 3.

Models	Dialogue lengths		
	< 300	300 − 600	> 600
HAN	63.34	53.61	40.06
PLA-HAN	63.55	54.43	41.62
AP-HAN	64.63	54.48	45.80

Table 3. Comparison of different lengths (Dataset 1)

As shown in Table 3: (1) The accuracy decreases with the increase of length, which indicates that it is with a greater challenge for long dialogue. (2) Our AP-HAN outperforms HAN and PLA-HAN in the dialogue of different lengths. Especially for long dialogue (>600), compared with HAN and PLA-HAN, the accuracy of intention classification increases significantly, reaching 5.74% and 4.18%, respectively.

Visualization of Attention. We further investigate the attention outputs of AP-HAN and HAN. An example dialogue (intent labeled as *consultation-business regulations*) is chosen from the test data for illustrating. We visualize the sentence attention in HAN and adjacency pair attention in AP-HAN in Figure 2. In this example, T4, T5, T8 and T13 are recognized as interrogative sentences. And according to the replication mechanism in Subsection 2.1, they are copied once because they appear in the second turn of an adjacent pair.

As shown in Figure 2, HAN model pays too much attention to the single turns (T2, T4, and T10) and makes a wrong

Turn	AP	AP-HAN	HAN	Dialogue
1	1	0.133	0.004	A: Hello It's a pleasure to serve you
2			0.550	B: Please check the phone traffic for me
3	2	0.105	0.074	A: Please wait a moment. Your data has been ran out this month
4			0.183	B: How much has it been excessively used so far
4	3	0.040	-	B: How much has it been excessively used so far **(copy)**
5			0.042	A: Ok, I'll check it for you
5	4	0.007	-	A: Ok, I'll check it for you **(copy)**
6			0.001	B: Ok
7	5	0.014	0.010	A: The cost at 24 o'clock last night was about twenty-one yuan and fifty-six cents more
8			0.015	B: It's more than 20 Yuan, isn't it
8	6	0.004	-	B: It's more than 20 Yuan, isn't it **(copy)**
9			0.001	A: Right
10	7	0.693	0.117	A: The data will be updated on the 1st. It will be available tomorrow
11			0.002	B: What time tomorrow
12	8	0.004	0.000	A: It' ll be 0 am.
13			0.001	B: Is it available after 0 o'clock tomorrow
13	9	0.000	-	B: Is it available after 0 o'clock tomorrow **(copy)**
14			0.000	A: Ok

Fig. 2. An example dialogue with utterance attention in HAN and adjacency pair attention in AP-HAN.

intent prediction of *consultation-account information*. In our model, the adjacency pair attention covers more complete important parts (T1-T4, T10-T11) which can capture more information about *consulting business regulations* and help the classifier make the right prediction in this example.

4. CONCLUSION

This paper focuses on incorporating dialogue structure into the hierarchical network model to learn a better dialogue representation and proposes an adjacency pair-aware hierarchical attention network (AP-HAN) for dialogue intent classification. Experimental results on public corpus show that the proposed model achieves significant improvements over the compared models, especially for long dialogues. Visualization of attention further illustrates the effectiveness of our model.

5. ACKNOWLEDGEMENTS

This work was supported by the Natural Science Foundation of Guangdong Province (2021A1515011864) and the National Natural Science Foundation of China (71472068).

6. REFERENCES

[1] P. Haffner, G. Tür, and J. Wright, "Optimizing SVMs for complex call classification," in *Proceedings of the 28th IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2003*, pp. 632–635.

[2] N. Kalchbrenner, E. Grefenstette, and P. Blunsom, "A convolutional neural network for modelling sentences," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, Volume 1: Long Papers*, pp. 655–665.

[3] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[4] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *3rd International Conference on Learning Representations, ICLR 2015*.

[5] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pp. 1746–1751.

[6] D. Tang, B. Qin, and T. Liu, "Document modeling with gated recurrent neural network for sentiment classification," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pp. 1422–1432.

[7] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, "Hierarchical attention networks for document classification," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2016*, pp. 1480–1489.

[8] S. Thornbury and D. Slade, *Conversation: From Description to Pedagogy*, Cambridge University Press, Cambridge, 2006.

[9] S. Mehri, E. Razumovskaia, T. Zhao, and M. Eskénazi, "Pretraining methods for dialog context representation learning," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pp. 3836–3845.

[10] Z. Zhang, J. Li, P. Zhu, H. Zhao, and G. Liu, "Modeling multi-turn conversation with deep utterance aggregation," in *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pp. 3740–3752.

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

[12] E. Schegloff and H. Sacks, "Opening up closing," *Semiotica*, vol. 8, no. 4, pp. 289–327, 1973.

[13] J. Li and G. Rao, "Formal classification and resource construction of Chinese questions," in *Proceedings of the 19th Chinese National Conference on Computational Linguistics, CCL 2020*, pp. 107–116.

[14] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Volume 1 (Long and Short Papers)*, pp. 4171–4186.

[15] K. Cho, B. Merrienboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pp. 1724–1734.

[16] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[17] J. Ding, P. Huang, J. Xu, and Y. Peng, "A human-machine dialogue intent classification method using utterance pseudo label attention," in *Proceedings of the 19th Chinese National Conference on Computational Linguistics, CCL 2020*, pp. 277–287.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015*.

[19] N. Vu, P. Gupta, H. Adel, and H. Schütze, "Bi-directional recurrent neural network with ranking loss for spoken language understanding," in *Proceedings of the 41st IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2016*, pp. 6060–6064.

[20] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *Proceedings of the 17th Annual Conference of the International Speech Communication Association, INTERSPEECH 2016*, pp. 685–689.