PHAED: A Speaker-Aware Parallel Hierarchical Attentive Encoder-Decoder Model for Multi-Turn Dialogue Generation

Zihao Wang, Ming Jiang, and Junli Wang

***Abstract*—This article presents a novel open-domain dialogue generation model emphasizing the differentiation of speakers in multi-turn conversations. Differing from prior work that treats the conversation history as a long text, we argue that capturing relative social relations among utterances (i.e., generated by either the same speaker or different persons) benefits the machine capturing fine-grained context information from a conversation history to improve context coherence in the generated response. Given that, we propose a Parallel Hierarchical Attentive Encoder-Decoder (PHAED) model that can effectively leverage conversation history by modeling each utterance with the awareness of its speaker and contextual associations with the same speaker's previous messages. Specifically, to distinguish the speaker roles over a multi-turn conversation (involving two speakers), we regard the utterances from one speaker as responses and those from the other as queries. After understanding queries via hierarchical encoder with inner-query and inter-query encodings, transformer-xl style decoder reuses the hidden states of previously generated responses to generate a new response. Our empirical results with three large-scale benchmarks show that PHAED significantly outperforms baseline models on both automatic and human evaluations. Furthermore, our ablation study shows that dialogue models with speaker tokens can generally decrease the possibility of generating non-coherent responses.**

***Index Terms*—Multi-turn dialogue generation, open-domain dialogue, speaker token, natural language generation, deep neural networks.**

I. Introduction

RECENTLY, substantive progress has been achieved in the field of natural language processing with the help of big textual data [1], [2], [3]. Dialogue generation is a text generation task in natural language processing, which aims to generate a reasonable response regarding a conversation history automatically. With the goal of providing AI-based virtual agents to support various services such as personal secretaries, companions to humans with emotional connections, and customer services, this task has become a popular research topic in academia and industry [4]. Traditional dialogue models are mainly task-oriented [5], [6], [7]. To improve the generalization ability of these AI models, recent studies have begun to focus on the development of open-domain conversational agents [8], [9], [10], [11].

The common practice of building a dialogue model is to train a sequence-to-sequence model using the conversation history to generate a context-coherent response. Considering the difficulty of understanding the complex scenarios in real life as social interaction, merchandising, and small talk, how to effectively understand conversation history is a critical challenge in the encoding process [12], [13]. To address this challenge, prior work usually proposes a learning-based model based on the framework of hierarchical recurrent encoder-decoder (HRED) [8], where the model contains word-level and utterance-level encoders to understand the conversation history.

Despite remarkable contributions made by prior work for capturing context information from the conversation history [9], [12], [13], [14], [15], [16], one major limitation of these studies is that they primarily focus on previous utterances' content but ignore the social relationships between these utterances (i.e., generated by the same or different speaker) [17]. We argue that such missing information is helpful to the machine in learning fine-grained context information by differentiating the content of previous utterances based on their speakers. By losing such fine-grained context information, the machine is difficult to capture the latent properties of the speaker (e.g., role) represented by the machine for response generation, and hence, hard to guarantee the context coherence of the generated response. As shown in Fig. 1, we can infer from the conversation history that *speaker0* knows the shoes' price from his message "These shoes are on sale". However, due to the lack of awareness of the speaker's information behind utterances, the response generated by an existing dialogue model based on transformer [18] mistakenly asks "How much is it".

The traditional model generates such erroneous responses ("How much is it") because it treats the conversation history as one long text, predicts a new response after the previous query, but fails to clearly distinguish between external

Manuscript received 13 January 2023; revised 1 August 2023; accepted 4 September 2023. Date of publication 18 September 2023; date of current version 12 January 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB4501700 and in part by the Shanghai Science and Technology Innovation Action Plan Project under Grant 22511100700. Recommended for acceptance by M. Huang. *(Corresponding author: Junli Wang.)*

Zihao Wang and Junli Wang are with the Key Laboratory of Embedded System and Service Computing (Tongji University), Ministry of Education, Shanghai 201804, China, and also with the National (Province-Ministry Joint) Collaborative Innovation Center for Financial Network Security, Tongji University, Shanghai 201804, China (e-mail: zhwang_tjuer@tongji.edu.cn; junliwang@tongji.edu.cn).

Ming Jiang is with the Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202 USA (e-mail: mj200@iu.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TBDATA.2023.3316472, provided by the authors.

Digital Object Identifier 10.1109/TBDATA.2023.3316472

2332-7790 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: MIT Academy Of Engineering. Downloaded on April 24,2025 at 07:31:20 UTC from IEEE Xplore. Restrictions apply.

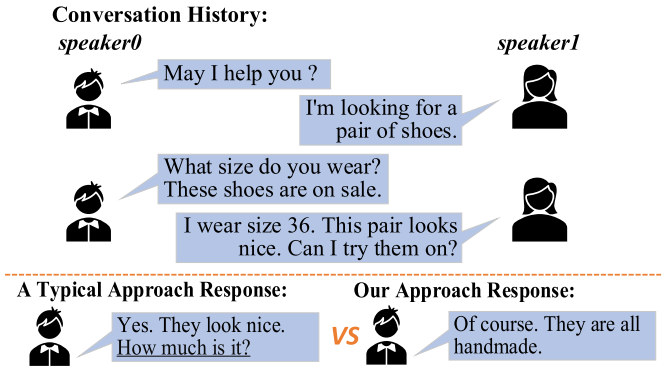


Fig. 1. Sample of the responses generated by a typical approach (Transformer) and our approach. *speaker0* and *speaker1* are speaker tokens. Underlined words indicate the part that does not meet the speaker token.

utterances (*speaker1*'s queries) and previously generated utterances (*speaker0*'s responses), resulting in the inconsistency between the new generated response ("How much is it") and previous responses ("May I help you", "These shoes are on sale").

To address the aforementioned limitation, we propose a speaker-aware learning model for improving multi-turn dialogue coherence by making full use of the conversation history of different speakers. The motivation of our work is to treat the queries and the responses from different speakers as two long hierarchical texts. Instead of mixing all utterances together as one long text, our approach aims to model each utterance with the awareness of its speaker and contextual associations with the same speaker's previous messages. Specifically, to make the model distinguish between queries and responses, we treat the conversational corpus as parallel data and add different speaker tokens at the beginning of queries and responses. First, a hierarchical encoder containing two-level encoding is proposed to obtain the local and global contextual representations of queries. Then, the decoder utilizes the turn-level recurrence and cross attention to take advantage of both previous responses and queries for generating the current response. In this way, since the responses are taken as one long coherent text in the decoder, the model tends to generate a new response that is consistent with the previously generated responses. Similarly, the queries are taken as another long coherent text and understood by the encoder. Moreover, we argue that it is unnecessary to re-understand past responses since the model must have understood a response before synthesizing it. Therefore, our decoder reuses hidden states of the previously generated responses instead of reconstructing these by the encoder. After considering the speaker roles, we can see from Fig. 1 that our approach generates a coherent response with respect to the context of *speaker0*.

Our main contributions include:

- We propose a novel dialogue generation model, PHAED, to generate context-coherent responses in multi-turn dialogues by dealing with utterance information with the awareness of their speakers.
- We design the hierarchical encoder and transformer-XL style decoder. Such design permits efficient utilization of both encoder and decoder hidden states that denote the information from the query speaker and responses speaker, respectively.
- By performing experiments on three public datasets, we show that our approach outperforms the strong baselines in terms of response quality. Besides, we conduct a fine-grained analysis of the performance of PHAED, which deepens our understanding of the characteristics of PHAED.

II. Related Work

A. Multi-Turn Dialogue Generation

As the multi-turn dialogue generation is accordant with the scenarios in daily life, it has gained increasing attention. Besides, since using plain texts has limitations for dialogue generation models, it is crucial to make the most of the semi-structured data (i.e., containing both textual and auctorial information).

Recent work on multi-turn dialogue generation mainly focuses on using conversation history effectively. Early, [8] successfully applies HRED [19] on dialogue generation that models conversation history via a hierarchical recurrent encoder and generates a response by a recurrent decoder. Most subsequent studies focus on designing exquisite attention mechanisms to detect the relevant words or utterances for response generation [9], [13], [15], [16], [20]. [15], [21], [22] use RNN-based variational auto-encoders to generate responses that are relevant to the content of the conversation history, but they do not take into account the differences between speakers. DialoGPT [20] further takes GPT2 as the basic structure to build the dialogue model. These working methods typically view dialogue context as a linear sequence of representations, and in recent years, some work has begun to focus on the overall structure of dialogue and coherence at the utterance level. To explore the discourse-level coherence among utterances, [23] proposes two training objectives. [24] uses an additional knowledge graph for constructing a dialogue graph, and requires the medical pre-trained model to serve as a supervision signal to train the model to recall pivotal information. [25] explores approaches to get latent discourse structures for dialogues based on pre-trained language models.

Compared with utterance-aware dialogue generation models that take dialogue history as a long text, both the content of the utterances and the speakers' roles are considered in the design of our model structure.

B. Speaker-Aware Dialogue Models

Regarding dialogue reasoning models, [26], [27], [28], [29], [30] prove that it is helpful to consider speaker roles in conversation. [26] proposes a speaker role-based contextual model for language understanding and dialogue policy learning. [27] proposes a speaker classification task in multi-party conversation. [28] studies the implicit discourse relation identification between different utterances. [29] finds that filling the gap between utterance-aware and speaker-aware representations benefits multi-turn response selection, but it is not applied to

Authorized licensed use limited to: MIT Academy Of Engineering. Downloaded on April 24,2025 at 07:31:20 UTC from IEEE Xplore. Restrictions apply.

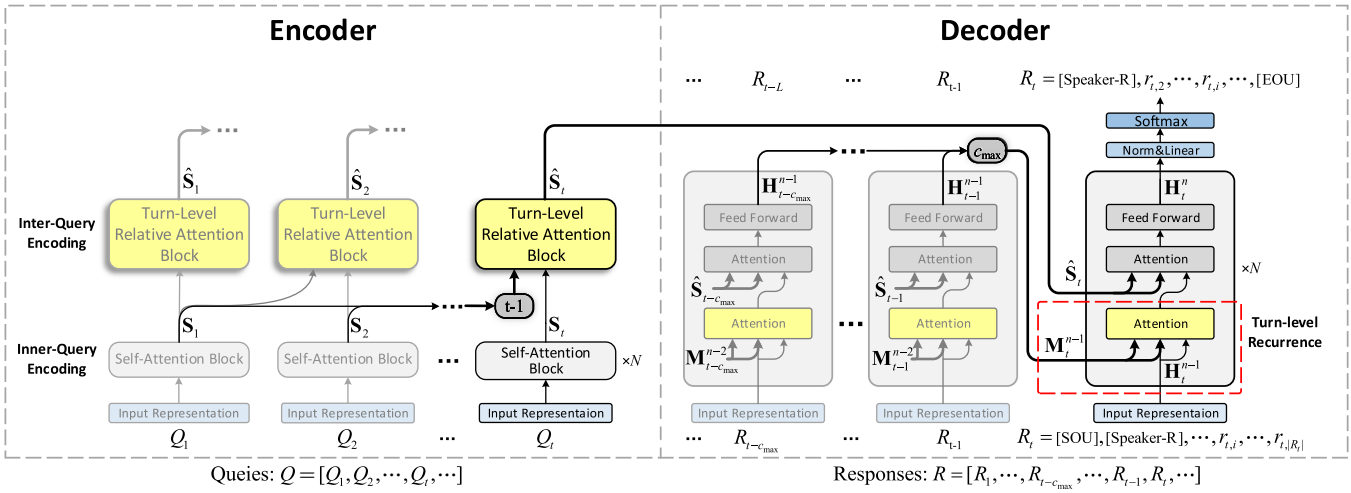


Fig. 2. Architecture of PHAED. Given a conversation involving a query set and a response set, $\mathbf{S}_t$ and $\hat{\mathbf{S}}_t$ denote the local and global contextual representations of the tth query Q_t. $\mathbf{H}_t^n$ denotes the hidden state of tth response from nth decoder block. c_{max} denotes the memory length of the decoder.

response generation. [30] proves that the main effect of speaker embeddings is important for dialogue-based relation extraction.

Regarding speaker-aware dialogue generation models, some models [31], [32], [33], [34], [35], [36], [37] require expensive extra personal information in addition to the current conversation. However, our goal is to leverage speakers' information from the current conversation [38], [39] even when there is no additional information. [31] considers extra fine-grained manual roles (*speaker* or *listener*) of each utterance. Some persona-based conversation models [32], [33], [34], [35] require a corpus containing specific identifiers of each person to extract the persona characteristics from the same person's conversations. However, for general conversational datasets [40], [41], [42] that are used in this article, in different conversations, speakers can only be unified and anonymous as speaker0 and speaker1, rather than labeling all speakers that occurred in the dataset with specific personal identifiers (e.g., names). In this general setting, [38] proposes a speaker-aware generative dialogue model that contains relative speaker embedding and relative utterance encoders. Besides, [11], [36], [39] use the pre-training model (e.g., GPT) and represent different speakers with different segment IDs.

Instead of requiring the extra manual personalized information [31], [37] or extra conversations containing specific personal identifiers [32], [33], [34], [35], [36], our approach is suitable for general conversational datasets [40], [41] without additional personal characteristics. Besides, differing from simply improving the data preprocessing by representing different speakers with different segments id [11], [36], [39], we design the appropriate model structure according to the difference of speakers.

III. Approach

Fig. 2 provides an overview of our approach. We take each conversation as parallel data. Following the regular dialog flow, regard utterances in each turn as a query-response pair, and the order of two speakers in a conversation is usually consistent [9], [41], [42]. With the assumption that differentiating the speaker of previous utterances should be helpful with the model's sensibility to the conversation context to generate coherent responses, our goal is to design a parallel multi-turn dialogue generation model (i.e., PHAED) with the consideration of the speaker role of utterances in the process of generating a context-coherent response in each turn.

Overall, we will describe PHAED from five aspects: 1) We first formalize the problem in Section III-A; 2) For modeling multi-turn conversation involving speaker roles, the input representation is designed in Section III; 3) Given queries from speaker-Q, the hierarchical attentive encoder constructs local contextual representations (inner-query encoding) and then combines all of them to obtain global contextual representations (inter-query encoding) in Section III-C; 4) After understanding queries, the decoder generates its current response based on the global contextual representations of the queries, the hidden states of its previous responses, and the local context of its partial current response in Section III-D. 5) In the end, we describe the model training process in Section III-E.

A. Preliminary

Suppose that we have a multi-turn dialogue dataset $\mathcal{D} = \{D^m\}_{m=1}^{M}$, where D^m is the mth conversation and M is the number of all conversations in the dataset. Each conversation D^m involves two speakers who give queries (i.e., `[Speaker-Q]`) and responses (i.e., `[Speaker-R]`) iteratively. Hence we represent each conversation D^m as a sequence of query-response pairs denoted as $D^m = \{(Q_t^m, R_t^m)\}_{t=1}^{T_m}$, where T_m is the number of pairs. Dialogue models often adopt an encoder-decoder framework. Here dialogue models aim to generate response R^m given queries Q^m in order. The training criterion is to maximize the conditional log-likelihood, which can be formulated as

$$\sum_{m=1}^{M}\sum_{t=1}^{T_m}\log P(R_t^m|Q_{\leqslant t}^m, R_{<t}^m;\theta),$$

Authorized licensed use limited to: MIT Academy Of Engineering. Downloaded on April 24,2025 at 07:31:20 UTC from IEEE Xplore. Restrictions apply.

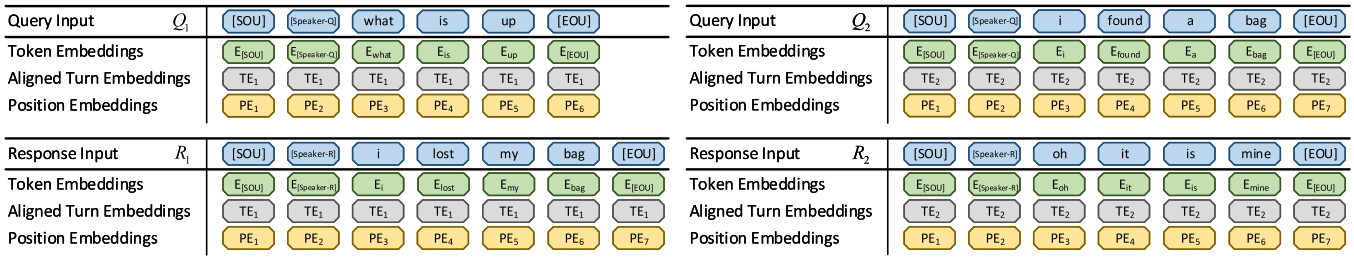


Fig. 3. Sample of the input representations.

where $Q^m_{\leqslant t}$ refers to all previous queries up to the tth turns, $R^m_{<t}$ denotes the previous responses prior to R^m_t, and θ denotes parameter of the model.

B. Input Representation

To distinguish the speaker identities over a multi-turn conversation, we design a novel speaker-aware input representation for words in the query and response utterances. More specifically, we first append two speaker tokens (i.e., `[Speaker-Q]` and `[Speaker-R]`) to the beginning of all queries and responses respectively. We then prepend a start-of-utterance token (i.e., `[SOU]`) and append an end-of-utterance token (i.e., `[EOU]`) to each utterance. Finally, we add the turn-level and token-level position embeddings to the token embedding as the input representation $\widetilde{q}_{t,i}$ and $\widetilde{r}_{t,i}$

$$\begin{aligned}\widetilde{q}_{t,i} &= E(q_{t,i}) + TE(t) + PE(i),\\ \widetilde{r}_{t,i} &= E(r_{t,i}) + TE(t) + PE(i),\end{aligned} \tag{1}$$

where $q_{t,i}$ ($r_{t,i}$) is the ith token in the tth query (response). $E(\cdot)$ looks up a token embedding from an embedding matrix. $TE(t)$ is the aligned turn-level position embedding indicating the tth utterance. $PE(i)$ is the token-level position embedding indicating the ith token in each utterance. All embeddings are learnable in training. A detailed visualization example of our input representation structure is provided in Fig. 3.

C. Hierarchical Encoder With Turn-Level Relative Attention

We want the encoder to capture and encode all the external information passed to PHAED. In other words, the encoder is responsible for understanding all queries from other people (i.e., speaker-Q). Since there are multiple queries from the same person and each query has its information, we use two steps to understand all queries. Fig. 2 (left) shows the architecture of our encoder. We first encode each query by self-attention blocks in *Inner-Query Encoding*. Then, we need to combine all information of queries, so we propose *turn-level relative attention* in *Inter-Query Encoding* to consider all queries comprehensively.

1) Inner-Query Encoding: To summarize the information from the individual query, we apply a standard N-layer Transformer encoder [18] to encode each query. Specifically, we obtain a representation matrix $\mathbf{S}^N_t \in \mathbb{R}^{|Q_t| \times d_s}$ for all tokens in the tth query Q_t from the top layer, where $|Q_t|$ is the length of Q_t. To simplify the notation, we skip the superscript N hereafter, i.e., $\mathbf{S}_t$. Notably, we adopt the pre-normalization [43] which has proven effective to stabilize the performance.

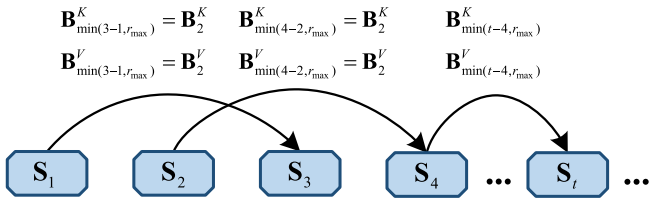


Fig. 4. Example edges referring to turn-level relative positions, or the turn-level distance between utterances. The vectors are learned for each turn-level relative position within a clipping maximum $r_{\max}$. The example assumes $2 \leq r_{\max}$. Note that not all edges are shown.

2) Inter-Query Encoding by Turn-Level Relative Attention: As historical contexts from the query speaker are crucial for understanding the current query, we aim to combine the information from all preceding queries to obtain a global context. To this end, we introduce an inter-query encoding method that obtains a global contextual representation $\hat{\mathbf{S}}_t \in \mathbb{R}^{|Q_t| \times d_s}$ for Q_t based on a set of the preceding queries denoted as $\mathcal{S}_{\leqslant t} = \{\mathbf{S}_1, \ldots, \mathbf{S}_t\}$, where each element is obtained in Section III-C1. Therefore, to combine the information from all the preceding queries, we propose a *turn-level relative attention* network that captures the dependencies of the queries using turn-level relative positions

$$\hat{\mathbf{S}}_t = \text{FFN}(\text{TurnRelAttn}(\mathbf{S}_t, \mathcal{S}_{\leqslant t}, \mathcal{S}_{\leqslant t})), \tag{2}$$

where $\text{FFN}(\cdot)$ denotes a feedforward network and $\text{TurnRelAttn}(\cdot)$ is our attention network that takes $\mathbf{S}_t$ as the query, and $\mathcal{S}_{\leqslant t}$ as the keys and values.

Turn-Level Relative Attention. As shown in Fig. 4, given the representation of each query, we propose the turn-level relative attention, which is extended from token-level [44] and segment-level [45], to model the global contextual representation by considering the turn-level relative position among queries in the history. So our model considers both absolute (aligned turn embeddings in input representation) and relative turn-level position information. In contrast, the existing dialogue model (e.g., ReCoSa [16]) only considers the absolute position information of utterances (e.g., TE_1, TE_2 in Fig. 3) when computing the global contextual representation And differing from segment-level relative attention used in machine translation [45], where each sentence attends to other sentences before and after itself, in our turn-level relative attention, the query can only

Authorized licensed use limited to: MIT Academy Of Engineering. Downloaded on April 24,2025 at 07:31:20 UTC from IEEE Xplore. Restrictions apply.

access the preceding queries because the queries are generated in order as the conversation progresses. Specifically, when doing an attention operation from the tth query to the past pth query, we first compute the attention's query, key, and value matrices by multiplying the corresponding weight matrices, i.e., $\mathbf{W}^Q, \mathbf{W}^K, \mathbf{W}^V \in \mathbb{R}^{d_s \times d_s}$, and then add the relative position information to the keys and values

$$\begin{aligned}\mathbf{Q}_t, \mathbf{K}_p, \mathbf{V}_p &= \mathbf{S}_t\mathbf{W}^Q, \mathbf{S}_p\mathbf{W}^K, \mathbf{S}_p\mathbf{W}^V, \forall \mathbf{S}_p \in \mathcal{S}_{\leqslant t} \\ \hat{\mathbf{K}}_p &= \mathbf{K}_p + (\mathbf{B}_r^K\mathbf{I})^\top, \\ \hat{\mathbf{V}}_p &= \mathbf{V}_p + (\mathbf{B}_r^V\mathbf{I})^\top, \\ r &= \min(t-p, r_{\max}),\end{aligned} \tag{3}$$

where r measures the relative position between t and p up to a pre-defined maximum number $r_{\max}$, $\mathbf{B}^K$, $\mathbf{B}^V \in \mathbb{R}^{d_s \times r_{\max}}$ are two learnable matrices that capture the relative position information for the attention's keys and values, and $\mathbf{I} \in \mathbb{R}^{1\times|Q_p|}$ is an all-one row vector. Here we take the rth column vector $\mathbf{B}_r^K$ from $\mathbf{B}^K$ and copy it $|Q_p|$ times across columns by multiplying it with $\mathbf{I}$. Similar operations apply to $\mathbf{B}_r^V$.

We then compute the attention matrix $\mathbf{A}_{t\to p} \in \mathbb{R}^{|Q_t|\times|Q_p|}$, and obtain a global contextual representation $\hat{\mathbf{S}}_t$ by a weighted sum over the values of all preceding queries, followed by a residual connection and a feedforward network as follows:

$$\begin{aligned}\mathbf{A}_{t\to p} &= \text{softmax}\left(\frac{\mathbf{Q}_t\hat{\mathbf{K}}_p^\top}{\sqrt{d_s}}\right), \\ \hat{\mathbf{S}}_t &= \text{FFN}\left(\mathbf{S}_t + \sum_{p=1}^{t}\mathbf{A}_{t\to p}\hat{\mathbf{V}}_p\right).\end{aligned} \tag{4}$$

D. Decoder With Turn-Level Recurrence

PHAED understands other people's queries through the encoder, but it still needs to consider its previously generated responses from speaker-R for generating its current response. Our idea is to store the hidden states of its previous responses as memory and reuse the memory as the information of these responses to generate its current response. For this purpose, as shown in Fig. 2 (right), we take the Transformer-xl [45], [46] with cross attention as the decoder.

Turn-Level Recurrence. In [46], the decoder consists of N Transformer layers, where each layer augments the attention's keys and values from the previous layer by caching the hidden states of a fixed length of previous words in the memory. This caching design allows the decoder to access a longer context in the memory. Similarly, we aim to extend the context from preceding responses for the decoder and encourage the decoder to capture the turn-level relationship between responses. To this end, we cache the hidden states of words from at most $c_{\max}$ previous responses and concatenate them along the length dimension as the memory $\mathbf{M}_t^{n-1}$ from the $(n-1)$th layer

$$\begin{aligned}c &= \max(t - c_{\max}, 1), \\ \mathbf{M}_t^{n-1} &= \text{SG}([\mathbf{H}_c^{n-1} \circ \cdots \circ \mathbf{H}_{t-2}^{n-1} \circ \mathbf{H}_{t-1}^{n-1}]),\end{aligned} \tag{5}$$

where $\circ$ stands for a concatenation operation. The function $\text{SG}(\cdot)$ denotes stop-gradient. $\mathbf{H}_c^{n-1} \in \mathbb{R}^{|R_c|\times d_s}$ in the memory denotes the hidden state of the cth response of word size $|R_c|$ from the $(n-1)$th transformer layer. Similar to [46], we truncate the gradient from the memory and augment the attention's keys and values with the memory. After that, the cross-attention module followed by a feedforward network fetches the queries context from encoder representation $\hat{\mathbf{S}}_t$ to obtain the next layer's hidden state $\mathbf{H}_t^n$

$$\begin{aligned}\widetilde{\mathbf{H}}_t^{n-1} &= [\text{SG}(\mathbf{M}_t^{n-1}) \circ \mathbf{H}_t^{n-1}], \\ \hat{\mathbf{H}}_t^n &= \text{Attention}(\mathbf{H}_t^{n-1}, \widetilde{\mathbf{H}}_t^{n-1}, \widetilde{\mathbf{H}}_t^{n-1}) \\ \mathbf{H}_t^n &= \text{FFN}(\text{Attention}(\hat{\mathbf{H}}_t^n, \hat{\mathbf{S}}_t, \hat{\mathbf{S}}_t)).\end{aligned} \tag{6}$$

Finally, we obtain the output probability of R_t by a linear layer and a softmax layer based on the final representation $\mathbf{H}_t^N$ from the top decoder layer

$$\begin{aligned}P(R_t|R_{<t}, Q_{\leqslant t}; \theta) &= \prod_{i=1}^{|R_t|} P(r_{t,i}|r_{t,<i}, R_{<t}, Q_{\leqslant t}; \theta) \\ &= \prod_{i=1}^{|R_t|} \text{softmax}(\mathbf{H}_{t,i}^N\mathbf{W}^O),\end{aligned} \tag{7}$$

where $\mathbf{W}^O \in \mathbb{R}^{d_s\times|V|}$ is a linear project matrix for a vocabulary of size $|V|$. $\mathbf{H}_{t,i}^N$ denotes the ith row vector from $\mathbf{H}_t^N$ and $|R_t|$ is the length of R_t.

E. Training Process

The proposed training process on dialogue data is illustrated in Algorithm 1, where the batch size is set to 1 for demonstration. In the practical training, we draw multiple dialogues as a mini-batch every iteration. As shown in Algorithm 1, for generating R_t, we reuse the hidden states of prior responses in the memory instead of reconstructing them by the encoder. In each iteration, we initialize $Memory$ to an empty set at the beginning and update model parameters at the end, ensuring stable process training. Besides, we stop the gradient of hidden states cached in $Memory$.

IV. Experiments

We conduct experiments on three datasets and compare PHAED with state-of-the-art baselines based on both automatic and human evaluations. In addition to showing the effectiveness of our method, we further conduct a fine-grained analysis to deepen our understanding of the characteristics of PHAED.

A. Experimental Setup

1) Datasets: Three popularly used benchmark datasets for open-domain multi-turn dialogue generation are adopted, which include:

- (1) *DailyDialog* [41]. Dialogues in DailyDialog cover various topics about our daily life, such as social activities and school life.

Authorized licensed use limited to: MIT Academy Of Engineering. Downloaded on April 24,2025 at 07:31:20 UTC from IEEE Xplore. Restrictions apply.

TABLE I
STATISTICS OF THREE DATASETS

	DailyDialog	PersonaChat	WoW	Ubuntu
train dialogues	11,118	8,939	18,430	1, 000, 000
valid dialogues	1,000	1,000	981	19,560
test dialogues	1,000	968	965	18,920
avg. utter. per dialogue	7.9	14.8	9.1	4.9
avg. tokens per utter.	14.6	12.9	16.6	16.2

Algorithm 1: Model Training Process.

Input: Untrained PHAED $\mathcal{M}(\boldsymbol{\theta})$, dialogue dataset $\mathcal{D}$
Output: Trained PHAED $\mathcal{M}(\boldsymbol{\theta})$
while not converge **do**
  Randomly Draw a dialogue from dataset: $D = \{(Q_t, R_t)\}_{t=1}^{T} \in \mathcal{D}$
  Initialize $Memory$ to an empty set
  for t from 1 to T do
    Get $\mathbf{S}_t$ with Q_t by Inner-Query Encoding in Section III-C1
    Get $\hat{\mathbf{S}}_t$ with $\{\mathbf{S}_t, \mathcal{S}_{\leqslant t}\}$ by Inter-Query Encoding in Section III-C2
    Initialize $\mathbf{H}_t^0$ to $[\widetilde{r}_{t,1}, \widetilde{r}_{t,2}, \ldots, \widetilde{r}_{t,|R_t|}]$ by (1)
    for n from 1 to N do
      Get $\mathbf{M}_t^{n-1}$ from $Memory$ by (5)
      Get $\mathbf{H}_t^n$ with $\{\mathrm{SG}\mathbf{M}_t^{n-1}, \mathbf{H}_t^{n-1}, \hat{\mathbf{S}}_t\}$ by (6)
      Update $Memory$ for caching the hidden states $\mathrm{SG}(\mathbf{H}_t^{n-1})$
    Compute the $P(R_t)$ by (7)
  Update parameters $\boldsymbol{\theta}$ by minimizing $-\sum_{t=1}^{T} logP(R_t)$
end while

- (2) *PersonaChat* [42]. PersonaChat is a knowledge-grounded dataset that contains dialogues and speaker profile information. Following the standard practice of prior work [12], we append the profile to the conversation history.
- (3) *WoW* [47]. WoW is an open-domain dialogue benchmark, where two annotators give each instance an initial topic. Its test set is categorized into Test Seen and Test Unseen based on whether the topic appears in the training set or not. We train our model and baselines on dialogues in the training set and evaluate all models on Test Seen.
- (4) *Ubuntu v2.0* [40]. In order to verify the effectiveness of the model on the large-scale dataset, we adopt the Ubuntu dataset,[1] which is a large multi-turn dialogue corpus extracted from Ubuntu question-answering forum.

We truncate the utterances with more than 50 tokens, and the truncated utterances with abnormal endings are corrected. Each Dialogue in all datasets involves two participants. Vocabulary is made up of all the words that appear throughout the data and are shared by queries and responses. Table I provides detailed statistics of each dataset.

[1][Online]. Available: https://github.com/rkadlec/ubuntu-ranking-dataset-creator

2) Compared Methods: We select several strong multi-turn dialogue generation models as baselines:

- (1) *HRAN*: A hierarchical recurrent encoder-decoder (HRED) model [8] equipped with hierarchical attention based on the utterance-level and the word-level presentations [9].
- (2) *DSHRED*: A HRED equipped with static and dynamic attention [15].
- (3) *SpkHRED*: A recent speaker-aware HRED with relative speaker embedding and relative utterance encoders [38].
- (4) *Transformer*: Under the encoder-decoder framework for dialogue generation, the most simple but natural idea is to directly use the Transformer [18] to encode all the previous utterances and then decode the representations to generate a response.
- (5) *ReCoSa*: A hierarchical transformer-based model that first encodes each utterance independently, and then combines all the information for response generation [16].
- (6) *DialoGPT*: Following [20], we train a multi-turn dialogue generation model on the basis of the GPT-2 [48]. Differing from [20], we train the baseline DialoGPT using the processed corpus with speaker tokens added.
- (7) *TransferTransfo*: A GPT-based speaker-aware dialogue model and the input representation is the sum of word embedding, speaker embedding, and position embedding [36], [39].
- (8) *DialogBERT*: A hierarchical Transformer architecture that captures the discourse-level coherence among utterances with two training objectives [23].

To compare the model structure fairly and eliminate the influence of data preprocessing, we train all models from scratch and use the same preprocessed data *containing speaker tokens* (except for TransferTransfo and DialogBERT). The main difference between DialoGPT and TransferTransfo is that DialoGPT uses the preprocessed data containing speaker tokens, while TransferTransfo uses segment id to represent different speakers.

3) Implementation Details: The dimension d_s of hidden states is set to be 512 in HRAN, DSHRED, Transformer(N = 6), ReCoSa(N = 6), PHAED(N = 4), and PHAED(N = 6). N denotes the number of layers. For DialoGPT(N=6 and N = 12) and SpkDialoGPT(N = 6), we use the small GPT-2 architecture with d_s 768. We increase the d_s to 560 in PHAED(N = 12) to make it has the same parameter size as DialoGPT(N = 12). Greedy search is taken as the decoding strategy. Adam optimizer [49] with an initial learning rate of 0.0005 is utilized for training, and the batch size is 32. We train and evaluate each model on a Tesla P100 card or V100 card with PyTorch. Our code is publicly available at: https://github.com/ZihaoW123/PHAED. Note that since [12] does not open their code and data, we use different preprocessed training/validation/test dialogues, baseline model hyper-parameters, and pre-trained vectors (for embedding-based metrics), resulting in different experimental results between us and [12].

4) Automatic Evaluation: We evaluate PHAED and baselines based on *coherence* and *diversity* metrics. For coherence evaluation, we adopt *Perplexity* [9], *BLUE-n* for n-grams (n = 1, 2, 3, 4) [13], and three embedding-based metrics [14].

Authorized licensed use limited to: MIT Academy Of Engineering. Downloaded on April 24,2025 at 07:31:20 UTC from IEEE Xplore. Restrictions apply.

TABLE II
AUTOMATIC EVALUATION RESULTS ON THREE DATASETS

Dataset	Model	Coherence								Diversity		$\|\theta\|$
		Perplexity ↓	BLEU-1 / 2 / 3 / 4 (%) ↑				Avg / Ext / Gre (%) ↑			Distinct-1 / 2 (%) ↑		
DailyDialog	HRAN	31.04	19.10	8.511	4.670	2.790	63.90	38.86	45.52	1.732	8.700	32.6M
	DSHRED	31.71	18.80	8.351	4.547	2.686	63.64	38.98	45.01	1.492	7.606	34.2M
	SpkHRED	34.22	19.21	8.421	4.485	2.533	63.84	38.51	45.11	1.052	4.510	40.2M
	PHAED(N=4)	**25.67**	**19.24**	**9.200**	**5.444**	**3.517**	**64.24**	**39.47**	**46.30**	**2.633**	**13.80**	42.7M
	Transformer(N=6)	27.34	17.72	7.205	3.722	2.092	62.85	37.88	44.58	2.398	11.85	50.2M
	ReCoSa(N=6)	25.34	17.77	7.186	3.684	2.100	62.90	37.42	44.89	2.481	12.39	68.5M
	DialoGPT(N=6)	29.93	17.69	8.528	5.214	3.564	64.12	**40.24**	45.98	1.905	12.37	57.2M
	TransferTransfo(N=6)	29.92	18.09	8.136	4.649	2.968	64.21	39.97	46.08	1.894	11.76	57.2M
	DialogBERT(N=6)	30.47	18.24	7.581	4.110	2.249	62.31	38.31	44.85	2.509	11.60	57.8M
	PHAED(N=6)	**24.45**	**19.02**	**9.174**	**5.508**	**3.602**	**64.42**	39.82	**46.34**	**2.932**	**15.58**	53.8M
	DialoGPT(N=12)	27.89	18.54	9.432	6.077	4.382	64.86	**40.70**	46.89	2.109	14.00	99.7M
	PHAED(N=12)	**23.71**	**20.47**	**10.61**	**7.089**	**5.326**	**64.91**	39.96	**47.34**	**3.639**	**20.19**	99.3M
PersonaChat	HRAN	36.59	21.06	10.22	5.292	2.779	62.18	38.21	43.48	0.2396	0.9633	34.1M
	DSHRED	36.96	**21.69**	10.46	5.401	2.841	62.58	38.56	44.17	0.2718	1.357	34.8M
	SpkHRED	38.21	21.19	10.08	5.088	2.580	61.95	37.97	43.37	0.1856	0.6902	40.7M
	PHAED(N=4)	**33.13**	21.48	**10.51**	**5.603**	**3.101**	**63.60**	**39.75**	**45.16**	**0.5401**	**2.580**	43.4M
	Transformer(N=6)	37.09	19.55	8.807	4.027	1.959	59.84	38.05	40.19	0.1443	0.4177	51.0M
	ReCoSa(N=6)	34.58	21.04	9.925	4.973	2.563	62.64	37.40	44.54	0.4417	1.703	69.4M
	DialoGPT(N=6)	32.91	20.89	10.19	5.369	2.958	62.69	39.38	44.43	**0.5126**	2.298	57.7M
	TransferTransfo(N=6)	32.69	20.72	10.04	5.189	2.728	61.04	39.06	42.88	0.4855	2.262	57.7M
	DialogBERT(N=6)	33.99	20.82	9.506	4.722	2.427	60.66	37.42	42.39	0.2024	1.533	58.6M
	PHAED(N=6)	**32.62**	**21.93**	**10.66**	**5.595**	**3.026**	**63.09**	**39.48**	**44.98**	0.4996	**2.453**	54.4M
WoW	Transformer(N=6)	52.56	17.54	5.872	2.320	1.079	65.46	33.84	46.29	1.345	4.812	78.1M
	ReCoSa(N=6)	48.20	18.42	7.267	3.196	1.608	66.47	34.19	46.73	1.762	7.655	80.1M
	DialoGPT(N=6)	42.91	19.30	7.761	3.653	1.956	66.27	37.17	47.65	2.047	11.25	78.0M
	TransferTransfo(N=6)	42.70	18.87	7.795	3.710	1.945	66.32	37.62	47.11	1.938	10.80	78.0M
	DialogBERT(N=6)	43.61	18.69	6.560	3.091	1.648	65.37	34.25	45.76	1.573	6.153	89.1M
	PHAED(N=6)	**39.94**	**19.41**	**7.871**	**3.772**	**2.056**	**66.61**	**38.45**	**48.27**	**2.553**	**12.78**	81.6M
Ubuntu	Transformer(N=6)	42.74	11.51	3.460	1.315	0.5546	**62.35**	34.08	43.64	0.08594	0.3306	82.9M
	ReCoSa(N=6)	36.40	12.74	4.750	2.183	1.150	59.39	34.65	43.05	0.3970	3.441	83.8M
	DialoGPT(N=6)	29.98	13.06	5.077	2.425	1.299	59.77	35.00	44.02	0.6783	5.425	81.7M
	TransferTransfo(N=6)	30.21	12.69	4.891	2.309	1.235	59.51	35.00	43.29	0.6713	5.141	81.7M
	PHAED(N=6)	**28.71**	**13.15**	**5.676**	**2.508**	**1.449**	60.43	**35.72**	**44.29**	**0.7328**	**5.735**	86.4M

All models use the same preprocessed data containing speaker tokens except for SpkDialoGPT, which represents different speakers with different segment id. N denotes the number of stacked layers. $|\theta|$ denotes the parameter size. Memory length c_{max} of PHAED is 1. The best results in each metric are highlighted with bold. "↑" means higher is better, and "↓" means lower is better.

Embedding-based metrics include Average(*Avg*), Extrema(*Ext*), and Greedy(*Gre*), using pre-trained Google news word embedding [50]. For diversity evaluation, we use *Distinct-1* and *Distinct-2*, which calculate the ratio of unique unigrams and bigrams. *Notably*, due to the speaker token (i.e., `[Speaker-R]`) being the first token in the generated response, the probability of the speaker token is much higher than that of other tokens, and the Perplexity with speaker token probability will be much lower than the Perplexity without that. Since we focus on other tokens in responses except for special tokens, *we adopt the Perplexity that does not include the probability of speaker token and other metrics evaluate the generated responses with speaker token removed.*

5) Human Evaluation: We further conduct a manual evaluation to explicitly examine the quality of dialogue models based on human judgments. Following prior work [22], [51], we randomly select 100 examples containing conversation history and responses generated by baselines and PHAED as testing examples. Based on such testing data, we recruit three human annotators (all graduate students) with good English skills to score the response quality on a scale of [0, 1, 2] from four aspects, which include: 1) *Fluency* in terms of the smoothness of response and the correctness of grammar, 2) *Coherence* indicating whether the response is coherent with conversation history, 3) *Informativeness* that focuses on the amount of information contained in the response, and 4) *Overall* that stands for the general evaluation, where 0, 1, and 2 indicate bad, good, and perfect responses, respectively.

B. Evaluation Results

Considering the influence of the parameter size, we compare PHAED($N = 4$) with RNN-based baselines and PHAED($N = 6$) with Transformer-based baselines. Table II shows the automatic evaluation results, where we observe that PHAED outperforms baselines in four datasets. For PHAED with different N, a small increase in N (from 4 to 6) makes PHAED perform better on Perplexity, and with substantial amplification of N (from 4 to 12 or 6 to 12), PHAED achieves better performance on all metrics. Taking the results on DailyDialog as an example, the metrics scores of the PHAED($N = 4$ and $N = 6$) are better than other baselines overall, and when the models' parameter sizes are the same, PHAED($N = 12$) outperforms DialoGPT($N = 12$) on most metrics. Therefore, we demonstrate that PHAED performs better than baselines on automatic evaluation and generates high-quality responses. There is no significant difference between the performance of DialoGPT and SpkDialoGPT, which also indicates that speaker tokens and speaker embedding make similar contributions to the model. Moreover, with respect to the lower value of Perplexity scores achieved by PHAED with larger N, we can infer that stacking more blocks benefits PHAED by increasing the possibility of generating coherent and diverse responses based on the conversation history.

We carry out the human evaluation on the DailyDialog that contains a wide variety of high-quality conversations about daily life [51]. As shown in Table III, PHAED achieves better performance than baselines on all the metrics. Our responses are fluent and coherent with the conversation context. Besides,

Authorized licensed use limited to: MIT Academy Of Engineering. Downloaded on April 24,2025 at 07:31:20 UTC from IEEE Xplore. Restrictions apply.

TABLE III
HUMAN EVALUATION RESULTS

Model	Flu.	Coh.	Inf.	Overall
HRAN	1.09	0.94	0.89	0.86
DSHRED	0.90	0.78	0.73	0.69
SpkHRED	0.89	0.77	0.69	0.66
Transformer	0.97	0.84	0.92	0.78
ReCoSa	1.00	0.84	0.86	0.77
DialogBERT	1.09	0.93	0.91	0.98
DialoGPT	1.15	0.97	0.92	1.04
TransferTransfo	1.18	0.99	0.90	0.95
PHAED	**1.28**	**1.19**	**1.21**	**1.14**

compared with the baselines, the scores of informativeness and overall are also higher, revealing that our responses are more informative and are preferred most by humans. The average Fleiss's kappa [52] is 0.42, indicating moderate agreement among the three annotators.

C. Ablation Study

To provide a fine-grained analysis of the contribution of each component in PHAED (i.e., speaker tokens and aligned turn embedding in input representations, turn-level relative attention in the encoder, and turn-level recurrence in the decoder), we conduct an ablation study. Table IV shows our results. The ablation models without speaker tokens show deteriorations in the majority of metrics such as perplexity, suggesting that simply adding speaker tokens benefits both PHAED and other dialogue models (i.e., Transformer and DialoGPT) by generating coherent and diverse responses regarding the conversation context. Without aligned turn embedding that encodes the order of utterances, PHAED achieves a decreasing performance in all metrics. Meanwhile, by removing turn-level relative attention block or memory relative attention, the performance also obviously decreases in all metrics. Therefore, it is critical to consider the utterance-level positional information and the contextual information of both queries and responses. Besides, PHAED without speaker tokens shows less deterioration than the other ablation models of PHAED in most metrics, suggesting that the components we proposed in PHAED play more important roles than the input representations with speaker tokens.

In the 'parallel' structure, PHAED models queries and responses separately. In order to sufficiently demonstrate the effectiveness of the proposed 'parallel' structure and turn-level relative attention in the encoder, we select several models with non-parallel structures for comparison in the ablation experiment: 1) HAED (Hierarchical Attentive Encoder-Decoder) with the same input representations as PHAED, which models both queries and responses in conversation history by our proposed hierarchical encoder with turn-level relative attention and generates the next response by a transformer decoder; 2) ReCoSa [16] that models both queries and responses with a proposed hierarchical encoder with self-attention and generates the next response by a transformer decoder; 3) Transformer [18] that models both queries and responses with transformer encoder and generates the next response by transformer decoder; 4) DialogGPT [48] that models both queries and responses by the GPT-style decoder.

As shown in Table IV (bottom), all metrics scores of PHAED significantly outperform HAED, which indicates that the model with a 'parallel' structure generates more coherent and diverse responses than the model with a traditional 'non-parallel' encoder-decoder structure. Similarly, PHAED also outperforms the decoder-only DialoGPT model on most metrics, indicating that the model with a 'parallel' structure is better than the 'non-parallel' decoder-only model. In addition, to prove the effectiveness of our proposed turn-level relative attention, we use HAED, ReCoSa, and Transformer for comparative experiments. Among them, the decoders of HAED, ReCoSa, and Transformer are the same, and the encoder of HAED uses the proposed turn-level relative attention, while the encoder of ReCoSa and Transformer uses self-attention. As shown in Table IV, HAED obtains better performance than ReCoSa and Transformer on most metrics, demonstrating that our proposed turn-level relative attention is superior to self-attention, and our encoder is superior to conversation history encoders of existing work.

By looking into the PHAED structure more in-depth, we further explore the impact of decoder memory length $L = c_{max}$ (i.e., the number of previously machine-generated responses cached in memory) on PHAED's performances to identify the optimal value of L. According to [46], the dependency length of the turn-level recurrence is the sum of the lengths of prior $N \cdot L$ responses. However, is L the bigger, the better? To answer this question, we re-train PHAED with different L. The values in the Fig. 5 represent the difference between the results of PHAED($L = 1$) (Table II) and the results of PHAED($L \geq 1$). Overall, with an increase of L, PHAED obtains a better perplexity, but there is only a small change in each metric score. Considering PHAED with a large memory length costs high computing resources, we empirically set the value of L as 3 in PHAED($N = 4$) and 2 in PHAED(N=6).

D. Case Study

We would like to know what PHAED has learned from the conversation history. We visualize the query-to-query weights of a conversation based on *turn-level relative attention* of PHAED(N=6). Formally, the weight of the tth query (row) attending to the pth query (column) is computed by $\alpha_t^p = \frac{1}{|Q_t|}\text{sum}(\mathbf{A}_{t\to p})$, where $\mathbf{A}_{t\to p}$ is defined by (4) and sum$(\cdot)$ gets the sum of all elements in the input matrix. As shown in Fig. 6, three findings are also common in other conversations: the attention scores of different utterances are really different, which proves that the importance of different utterances is different, so it is necessary for the model to recognize different utterances by the proposed turn-level attention; the first query Q_1 (first column), which contains the major topic of a conversation, seems to be a terrifically useful context for subsequent queries; since the hidden states of current query can be passed by the residual connection, the turn-level attention appears to care less about the current query (diagonal queries).

Furthermore, we show two examples from the DailyDialog test result in Table V. In example 1, PHAED generates an

Authorized licensed use limited to: MIT Academy Of Engineering. Downloaded on April 24,2025 at 07:31:20 UTC from IEEE Xplore. Restrictions apply.

TABLE IV
ABLATION STUDY RESULTS(%). WE ALSO EVALUATE TEST RESULTS WITH THE SPEAKER TOKENS REMOVED

Model	Coherence								Diversity	
	Perplexity ↓	BLEU-1 / 2 / 3 / 4 ↑				Avg / Ext / Gre ↑			Distinct-1 / 2 ↑	
PHAED (N=4)	**25.67**	**19.24**	**9.200**	**5.444**	**3.517**	**64.24**	39.47	**46.30**	**2.633**	**13.80**
w/o speaker tokens	25.77	18.05	8.416	4.855	3.045	63.86	**39.57**	45.58	2.364	12.73
w/o aligned turn embedding	26.03	18.20	8.154	4.461	2.636	63.73	39.28	45.49	2.339	12.40
w/o turn-level relative attention	26.15	17.37	7.905	4.492	2.772	62.87	38.88	44.81	2.335	12.08
w/o turn-level recurrence	25.99	18.42	8.264	4.626	2.815	63.72	39.27	45.72	2.249	11.70
PHAED (modeling Q by our encoder and modeling R by our decoder, N=6)	**24.45**	**19.02**	**9.174**	**5.508**	**3.602**	**64.42**	39.82	**46.34**	**2.932**	**15.58**
HAED (modeling Q and R by our hierarchical encoder with turn-level relative attention, N=6)	25.06	18.11	7.409	3.866	2.263	62.91	37.31	44.72	2.629	13.37
ReCoSa (modeling Q and R by a hierarchical encoder with self-attention, N=6)	25.34	17.77	7.186	3.684	2.100	62.90	37.42	44.89	2.481	12.39
Transformer (modeling Q and R by traditional transformer, N=6)	27.34	17.72	7.205	3.722	2.092	62.85	37.88	44.58	2.398	11.85
w/o speaker tokens	27.93	17.48	7.078	3.620	1.966	62.51	37.78	44.44	1.967	9.483
DialoGPT (modeling Q and R by GPT decoder, N=6)	29.93	17.69	8.528	5.214	3.564	64.12	**40.24**	45.98	1.905	12.37
w/o speaker tokens	30.86	17.71	8.474	5.124	3.458	63.91	39.90	45.78	1.869	11.96

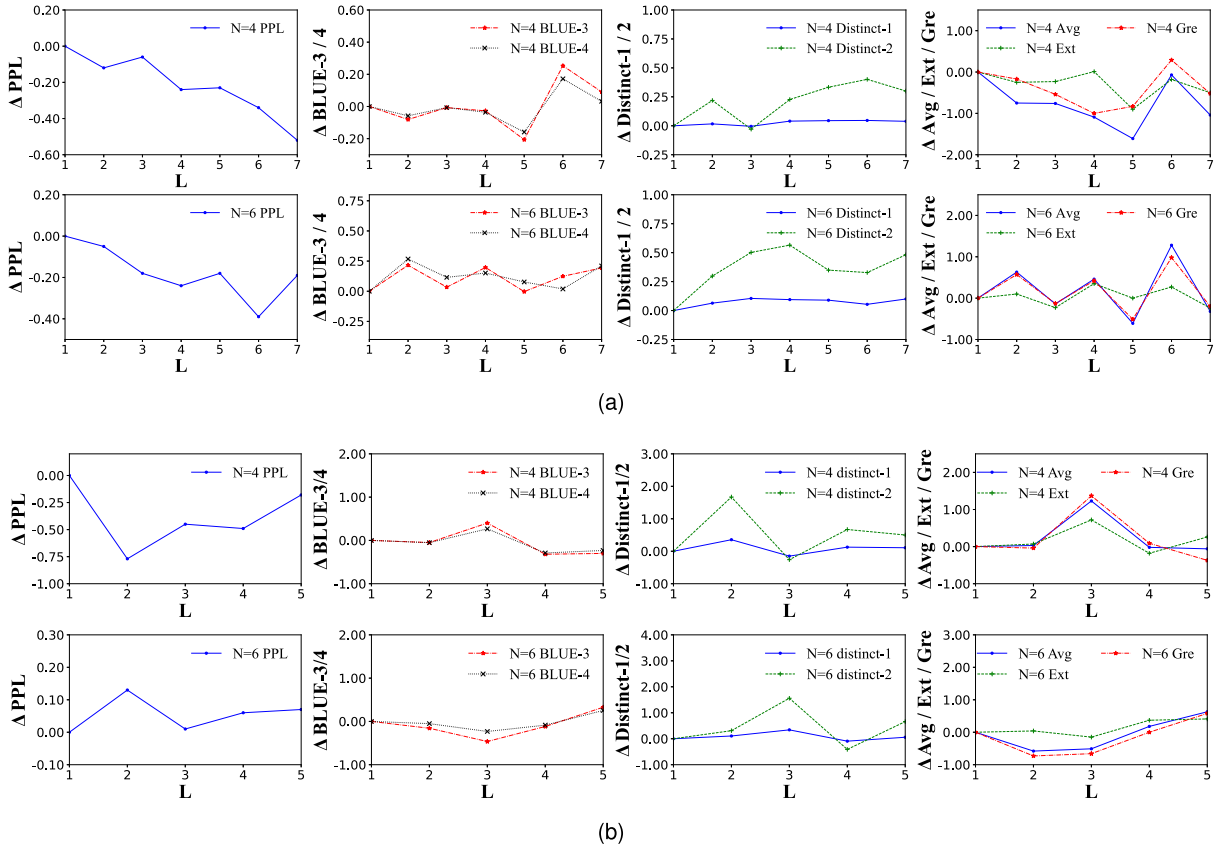


Fig. 5. Impact of memory length $L = c_{max}$ on the performance of PHAED on PersonaChat (a) and DailyDialog, (b) **Δ** *metric* represents the change in the score of the metric. Perplexity is abbreviated as *PPL*.

appropriate and informative response, but other baselines either generate responses from the wrong speaker perspective or generate short and safe responses. In example 2, PHAED generates a response that includes clear location information. However, DialGPT and ReCoSa generate responses based on a wrong previous query, and the other responses only contain fuzzy location information. We also include a case study with chatgpt [53] in Appendix A, available online, and we find that chatgpt is stronger than our model in terms of diversity, but we have to design different prompts for different scenarios to ensure coherence.

E. Discussion on Utilizing the Large-Language Models

Nowadays, LLM-based dialogue models [53] usually use pre-trained large language models to fine-tune dialogue data. In this process, although the model can learn to talk through imitation, its process of learning dialogue is not the same as that of humans.

We think that if the model is regarded as an independent person, the encoder is used to understand all external information, and the decoder is responsible for generating utterances that conform to its identity based on the information it understands. That is, the human-centered world is divided into external and internal worlds, the encoder understands the external objective

Authorized licensed use limited to: MIT Academy Of Engineering. Downloaded on April 24,2025 at 07:31:20 UTC from IEEE Xplore. Restrictions apply.

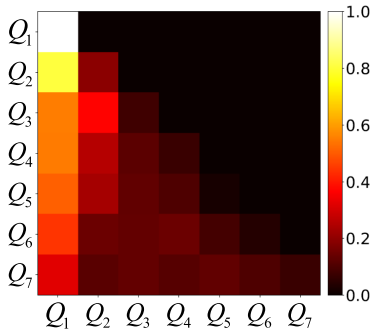


Fig. 6. Visualization of weights from query-to-query. The total weights of each row are 1.

TABLE V
TWO SAMPLES FROM THE DAILYDIALOG TEST RESULT

Sample 1	
History	Speaker-R May i help you? Speaker-Q I'm looking for a pair of shoes. Speaker-R What size do you wear? These shoes are on sale. Speaker-Q I wear size 36. This pair looks nice. Can I try them on?
HRAN	Speaker-R Of course.
DSHRED	Speaker-R Of course. It's a bit too expensive.
SpkHRED	Speaker-R Of course. You can have a look at the moment.
Transformer	Speaker-R Yes. They look nice. How much is it?
ReCoSa	Speaker-R Yes, they are.
DialGPT	Speaker-R Yes, of course.
TransferTransfo	Speaker-R I think they look too tight.
PHAED	Speaker-R Of course. They are all handmade.
Sample 2	
History	Speaker-Q 911. How may I help you? Speaker-R I need to report a break in. Speaker-Q When did this break in occur? Speaker-R Last night. Speaker-Q Where did the break in happen?
HRAN	Speaker-R I was in the middle of the city.
DSHRED	Speaker-R It was in the middle of the city.
SpkHRED	Speaker-R Well, I'm sorry, sir. We have a nice room.
Transformer	Speaker-R It was a very good game.
ReCoSa	Speaker-R It was in the middle of October.
DialGPT	Speaker-R It's a couple of days ago.
TransferTransfo	Speaker-R What do you mean?
PHAED	Speaker-R It happened at the hilton hotel.

world, and the decoder represents the internal world to transmit information to the outside world. So compared with the LLM-based dialogue model, the training process of our model is more in line with the process of human learning to speak. Besides, since our model decouples the two functions of understanding and generation, the encoder and decoder of our model can be initialized with existing "encoder-decoder" style LLM (e.g., T5 [54] and BART [55]). Therefore, we believe that there is still room for such a method when taking into account the advantages of LLM-based methods in improving response quality.

In the future, we will take advantage of the LLM and extend PHAED in multi-party conversation that includes multiple speakers, where the decoder generates the utterances of the self-speaker (e.g., `Speaker0`), and the encoder is responsible for understanding the utterances of all other speakers (e.g., `Speaker1`, `Speaker2`, `Speaker3`, ...).

V. CONCLUSION

We have presented a novel learning model called PHAED for multi-turn dialogue generation by utilizing utterance relations based on their speakers to capture fine-grained conversation context information. Unlike prior methods that mainly focus on the conversation history content to design an appropriate model structure, we design the hierarchical encoder and transformer-XL style decoder that emphasize the contextual associations of the same speaker's utterances for response generation. Such design permits efficient utilization of both encoder and decoder hidden states. Such design permits efficient utilization of both encoder and decoder hidden states. The presented experiments with three benchmark datasets have shown that PHAED outperforms the strong baselines in terms of response quality on both automatic and human evaluations. Moreover, we find that PHAED learns more from utterances containing high-level topic information of a conversation history than other utterances.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. N. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 4171–4186.

[2] Q. Li, M. Cheng, J. Wang, and B. Sun, "LSTM based phishing detection for big Email data," *IEEE Trans. Big Data*, vol. 8, no. 1, pp. 278–288, Feb. 2022.

[3] H. Poostchi and M. Piccardi, "BiLSTM-SSVM: Training the BiLSTM with a structured hinge loss for named-entity recognition," *IEEE Trans. Big Data*, vol. 8, no. 1, pp. 203–212, Feb. 2022.

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

[5] M. Henderson, B. Thomson, and S. Young, "Deep neural network approach for the dialog state tracking challenge," in *Proc. SIGDIAL Conf.*, Metz, France: Association for Computational Linguistics, 2013, pp. 467–471. [Online]. Available: https://www.aclweb.org/anthology/W13--4073

[6] T. Zhao and M. Eskenazi, "Towards end-to-end learning for dialog state tracking and management using deep reinforcement learning," in *Proc. 17th Annu. Meeting Special Int. Group Discourse Dialogue*, Los Angeles: Association for Computational Linguistics, 2016, pp. 1–10. [Online]. Available: https://www.aclweb.org/anthology/W16-3601

[7] A. Madotto, C.-S. Wu, and P. Fung, "Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, Australia:Association for Computational Linguistics, 2018, pp. 1468–1478. [Online]. Available: https://www.aclweb.org/anthology/P18-1136

[8] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 3776–3783.

[9] C. Xing, W. Wu, Y. Wu, M. Zhou, Y. Huang, and W.-Y. Ma, "Hierarchical recurrent attention network for response generation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, Art. no. 688.

[10] C. Chen, J. Peng, F. Wang, J. Xu, and H. Wu, "Generating multiple diverse responses with multi-mapping and posterior mapping selection," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, 2019, pp. 4918–4924. [Online]. Available: https://doi.org/10.24963/ijcai.2019/683

[11] S. Bao, H. He, F. Wang, H. Wu, and H. Wang, "PLATO: Pre-trained dialogue generation model with discrete latent variable," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 85–96.

Authorized licensed use limited to: MIT Academy Of Engineering. Downloaded on April 24,2025 at 07:31:20 UTC from IEEE Xplore. Restrictions apply.

[12] Y. Zhao, C. Xu, and W. Wu, "Learning a simple and effective model for multi-turn response generation with auxiliary tasks," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, Association for Computational Linguistics, 2020, pp. 3472–3483. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-main.279

[13] Z. Tian, R. Yan, L. Mou, Y. Song, Y. Feng, and D. Zhao, "How to make context more useful? An empirical study on context-aware neural conversational models," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 231–236.

[14] I. V. Serban et al., "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 3295–3301.

[15] W. Zhang et al., "A static and dynamic attention framework for multi turn dialogue generation," *ACM Trans. Inf. Syst.*, vol. 41, no. 1, Jan. 2023, Art. no. 15. [Online]. Available: https://doi.org/10.1145/3522763

[16] H. Zhang, Y. Lan, L. Pang, J. Guo, and X. Cheng, "ReCoSa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3721–3730.

[17] D. Hovy and D. Yang, "The importance of modeling social factors of language: Theory and practice," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2021, pp. 588–602.

[18] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[19] A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, and J.-Y. Nie, "A hierarchical recurrent encoder-decoder for generative context-aware query suggestion," in *Proc. 24th ACM Int. Conf. Inf. Knowl. Manage.*, 2015, pp. 553–562.

[20] Y. Zhang et al., "DIALOGPT : Large-scale generative pre-training for conversational response generation," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics: Syst. Demonstrations*, 2020, pp. 270–278.

[21] B. Sun, S. Feng, Y. Li, J. Liu, and K. Li, "Generating relevant and coherent dialogue responses using self-separated conditional variational AutoEncoders," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, Association for Computational Linguistics, 2021, pp. 5624–5637. [Online]. Available: https://aclanthology.org/2021.acl-long.437

[22] J. Xu, Z. Lei, H. Wang, Z.-Y. Niu, H. Wu, and W. Che, "Discovering dialog structure graph for coherent dialog generation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, Association for Computational Linguistics, 2021, pp. 1726–1739. [Online]. Available: https://aclanthology.org/2021.acl-long.136

[23] X. Gu, K. M. Yoo, and J.-W. Ha, "DialogBERT: Discourse-aware response generation via learning to recover and rank utterances," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 12911–12919. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/17527

[24] Y. Zhao et al., "Medical dialogue response generation with pivotal information recalling," in *Proc. 28th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, New York, NY, USA, 2022, pp. 4763–4771. [Online]. Available: https://doi.org/10.1145/3534678.3542674

[25] C. Li, P. Huber, W. Xiao, M. Amblard, C. Braud, and G. Carenini, "Discourse structure extraction from pre-trained and fine-tuned language models in dialogues," in *Proc. Findings Assoc. Comput. Linguistics*, Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 2562–2579. [Online]. Available: https://aclanthology.org/2023.findings-eacl.194

[26] T.-C. Chi, P. C. Chen, S.-Y. Su, and Y.-N. Chen, "Speaker role contextual modeling for language understanding and dialogue policy learning," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, 2017, pp. 163–168.

[27] Z. Meng, L. Mou, and Z. Jin, "Towards neural speaker modeling in multi-party conversation: The task, dataset, and models," in *Proc. 11th Int. Conf. Lang. Resour. Eval.*, Miyazaki, Japan, 2018, pp. 3142–3145. [Online]. Available: https://www.aclweb.org/anthology/L18--1496

[28] M. D. Ma, K. Bowden, J. Wu, W. Cui, and M. Walker, "Implicit discourse relation identification for open-domain dialogues," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 666–672.

[29] L. Liu, Z. Zhang, H. Zhao, X. Zhou, and X. Zhou, "Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue," 2020, *arXiv:2009.06504*.

[30] M. Zhou, D. Ji, and F. Li, "Relation extraction in dialogues: A deep learning model based on the generality and specialty of dialogue text," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 2015–2026, 2021.

[31] Y. Liu, H. Qian, H. Xu, and J. Wei, "Speaker or listener? The role of a dialogue agent," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 4861–4869.

[32] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and B. Dolan, "A persona-based neural conversation model," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, Berlin, Germany: Association for Computational Linguistics, 2016, pp. 994–1003. [Online]. Available: https://www.aclweb.org/anthology/P16--1094

[33] O. O. Olabiyi, A. Khazane, and E. T. Mueller, "A persona-based multi-turn conversation model in an adversarial learning framework," in *Proc. 17th IEEE Int. Conf. Mach. Learn. Appl.*, 2018, pp. 489–494.

[34] J. Bak and A. Oh, "Variational hierarchical user-based conversation model," in *Proc. Conf. Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 1941–1950. [Online]. Available: https://www.aclweb.org/anthology/D19--1202

[35] Z. Chan et al., "Modeling personalization in continuous space for response generation via augmented Wasserstein autoencoders," in *Proc. Conf. Empir. Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, Hong Kong, China: Association for Computational Linguistics, 2019, pp. 1931–1940. [Online]. Available: https://www.aclweb.org/anthology/D19-1201

[36] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "TransferTransfo: A transfer learning approach for neural network based conversational agents," 2019, *arXiv:1901.08149*.

[37] Y. Zheng, R. Zhang, M. Huang, and X. Mao, "A pre-training based personalized dialogue generation model with persona-sparse data," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 9693–9700.

[38] T. Zhao and T. Kawahara, "Effective incorporation of speaker information in utterance encoding in dialog," 2019, *arXiv:1907.05599*.

[39] Y. Wang et al., "A large-scale chinese short-text conversation dataset," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, X. Zhu, M. Zhang, Y. Hong, and R. He, Eds., Cham: Springer International Publishing, 2020, pp. 91–103.

[40] R. Lowe, N. Pow, I. Serban, and J. Pineau, "The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems," in *Proc. 16th Annu. Meeting Special Int. Group Discourse Dialogue*, 2015, pp. 285–294.

[41] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "DailyDialog: A manually labelled multi-turn dialogue dataset," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, 2017, pp. 986–995.

[42] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, 2018, pp. 2204–2213.

[43] S. Bao et al., "PLATO-2: Towards building an open-domain chatbot via curriculum learning," 2020, *arXiv:2006.16779*.

[44] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2018, pp. 464–468.

[45] Z. Zheng, X. Yue, S. Huang, J. Chen, and A. Birch, "Towards making the most of context in neural machine translation," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, C. Bessiere, Ed., 2020, pp. 3983–3989. [Online]. Available: https://doi.org/10.24963/ijcai.2020/551

[46] Z. Dai, Z. Yang, Y. Yang, J. G. Carbonell, Q. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2978–2988.

[47] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents," in *Proc. Int. Conf. Learn. Representations*, 2019.

[48] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," Tech. Rep., OpenAI, 2019.

[49] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015.

[50] T. Mikolov, G. Corrado, C. Kai, and J. Dean, "Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Representations*, 2013.

[51] H. Cai, H. Chen, Y. Song, C. Zhang, X. Zhao, and D. Yin, "Data manipulation: Towards effective instance learning for neural dialogue generation via learning to augment and reweight," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Association for Computational Linguistics, 2020, pp. 6334–6343. [Online]. Available: https://www.aclweb.org/anthology/2020.acl-main.564

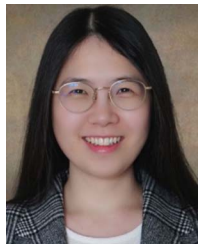
[52] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, pp. 378–382, 1971.

[53] OpenAI, "ChatGPT," 2023. [Online]. Available: https://openai.com/blog/chatgpt/

Authorized licensed use limited to: MIT Academy Of Engineering. Downloaded on April 24,2025 at 07:31:20 UTC from IEEE Xplore. Restrictions apply.

[54] C. Raffel et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," 2019, *arXiv:1910.10683*. [Online]. Available: http://arxiv.org/abs/1910.10683
[55] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Association for Computational Linguistics, 2020, pp. 7871–7880. [Online]. Available: https://aclanthology.org/2020.acl-main.703

Zihao Wang is currently working toward the PhD degree in computer science and technology with Tongji University, and his main research interest is open-domain dialog.



Ming Jiang received the PhD degree in informatics from the University of Illinois Urbana-Champaign. She is an assistant professor in data science with the Luddy School of Informatics, Computing, and Engineering, Indiana University-Purdue University Indianapolis. Her research broadly focuses on trustworthy natural language processing and artificial intelligence.

Junli Wang received the PhD degree in computer science from Tongji University, Shanghai, China, in 2007. She is currently an associate researcher with the College of Electronics and Information Engineering, Tongji University. Her research interests include text data analysis, deep learning, and artificial intelligence.

Authorized licensed use limited to: MIT Academy Of Engineering. Downloaded on April 24,2025 at 07:31:20 UTC from IEEE Xplore. Restrictions apply.