# Local and Global Contexts for Conversation

**Zuoquan Lin** and **Xinyi Shen**
Information and Computation Science Department
Peking University, Beijing, China
{linzuoquan,xinyi.shen}@pku.edu.cn

## Abstract

The context in conversation is the dialog history crucial for multi-turn dialogue. Learning from the relevant contexts in dialog history for grounded conversation is a challenging problem. Local context is the most neighbor and more sensitive to the subsequent response, and global context is relevant to a whole conversation far beyond neighboring utterances. Currently, pretrained transformer models for conversation challenge capturing the correlation and connection between local and global contexts. We introduce a *local and global conversation model* (LGCM) for general-purpose conversation in open domain. It is a local-global hierarchical transformer model that excels at accurately discerning and assimilating the relevant contexts necessary for generating responses. It employs a local encoder to grasp the local context at the level of individual utterances and a global encoder to understand the broader context at the dialogue level. The seamless fusion of these locally and globally contextualized encodings ensures a comprehensive comprehension of the conversation. Experiments on popular datasets show that LGCM outperforms the existing conversation models on the performance of automatic metrics with significant margins.[1]

## 1 Introduction

The role of context is significant in the similarity of words in a language. The contexts of a word are the neighboring tokens or grammatical structures. Contextualized embeddings encode both words and their contexts and generate contextualized representations. Language modeling captures distributed semantics embedded within these contextualized representations. The transformer-based pretrained language models (LMs) have become a foundation for NLP-like tasks (Bommasani et al., 2021). A well-established best practice in the field has consistently demonstrated that the utilization of large language models (LLMs) tends to yield superior performance in a wide range of NLP tasks, including conversational applications (say (Wolf et al., 2019; Adiwardana et al., 2020; Roller et al., 2021; Reed et al., 2022; Thoppilan et al., 2022), among others).

*Conversation models* (CMs) are generative sequence-sequence models for general-purpose conversations and learn the multi-agent distribution of utterances simultaneously. Most existing CMs are based on LMs, in which the LMs are used for accomplishing conversation by collaboration between agents that own their LMs or share a single LM in the spirit of parameter sharing (PS), where multiple models share the parameters in part or whole. In this paper, we consider the CMs with a single LM for two-agent conversation, such as human-machine dyadic dialogue.

More specifically, CMs use either vanilla Transformer (Vaswani et al., 2017) as single-turn dialogue, such as question answering, where only the current utterance is considered as the history at any given turn, or for multi-turn dialogue adapt the Transformer architecture by concatenating multiple turns sequentially to capture the evolving context (Wolf et al., 2019; Oluwatobi and Mueller, 2020; Zhang et al., 2019a). Prominent examples of such CMs include TransferTransfo (Wolf et al., 2019), Meena (Adiwardana et al., 2020), Blender (Roller et al., 2021), Athena (Reed et al., 2022) and LaMDA (Thoppilan et al., 2022), among others.

The context in conversation is the dialog history crucial for multi-turn dialogue. CMs require an understanding of the dialog history, in the context of previous pairwise utterances and the current query at any turn. For example, as humans in everyday dialogue, the speaker's intent often cannot be detected by looking at the utterance level. In contrast, the speaker's acts are specific to each utterance and

[1]Our codes are available at https://github.com/PKUAI-LINGroup/LGCM.



*Findings of the Association for Computational Linguistics: EACL 2024*, pages 1408–1418
March 17-22, 2024 

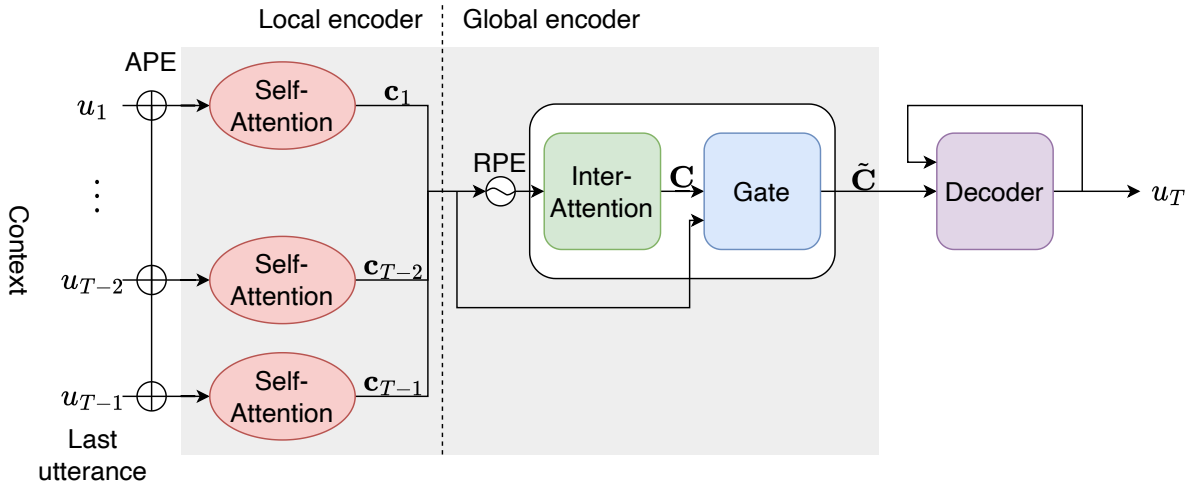



Figure 1: The architecture of LGCM: The encoder is hierarchical attention consisting of the local and global encoders. The local encoders are standard transformer modules with PS (depicted the same color as Self-Attention) for each utterance in context. The global encoder consists of Inter-Attention and Gate for contextualized representations, which are sent to the cross-attention in the decoder. The decoder is a standard transformer decoder.

change throughout a whole dialog history at the dialogue level. One of the key challenges faced by CMs lies in striking the right balance between staying current which involves giving preference to recent utterances, and drawing from the past effectively accumulating a prior understanding of the dialogue. The process of learning the relevant historical contexts necessary for fostering grounded and meaningful conversations remains a challenging problem in this domain.

A criticism of the existing CMs is their inability to effectively utilize the available dialog history and gain a comprehensive view of a conversation (Sankar et al., 2019). A common problem of those CMs is their failure to establish meaningful correlations and connections between individual utterances. They often treat all the words as a single sequence and concatenate multiple turns in history into a single sequence, which neglects the distinct contexts of individual utterances within the broader dialogue history.

To address the inherent problem of current CMs, we propose a more nuanced approach. In our model, we define each utterance as *local context* for tokens at the utterance level and whole a dialogue as *global context* for inter-utterances at the dialogue level. Moreover, we find it valuable to position the relationships among inter-utterances within a dialog history relative to one another. In our model, the conversation at different turns tells on each other, and all together, they tell what we talk about.

Namely, we introduce a *local and global CM* (LGCM) for multi-turn dialogue in open domain. It is a local-global hierarchical transformer model, illustrated in Figure 1. It is an encoder-decoder architecture in which the decoder is the same as Transformer (Vaswani et al., 2017) with the cross-attention between the encoder and the decoder, but the encoder is a hierarchical attention structure. The encoder of LGCM consists of *local encoders* and *global encoder*. The local encoders are implemented by a standard transformer module (Self-Attention) for each utterance in the local context using absolute position encoding (APE). The global encoder consists of *Inter-Attention* and *Gate* for contextualized representations in the global context, which are sent to the cross-attention in the decoder. The inter-attention is the attention between the current and all the utterances using relative positional encoding (RPE) (Shaw et al., 2018). The gate fuses the representations of the local encoders and the inter-attention by a nonlinear transformation for local-global contextualized representation, see explanation in the subsection 3.2.

In summary, the main contributions of this paper are the following:

(1) We are first trying to propose a CM that makes the connections between local context at the utterance level and global context at the dialogue level in a coherent way.

(2) We propose a new attention mechanism (Inter-Attention) between current and historic utter-

ances using RPE, which can separately deal with each utterance in a context. We extend the RPE from a single sequence in the self-attention to pairwise utterances within the conversation.

Experiments on popular datasets (DailyDialog, MultiWOZ, PersonaChat) show that LGCM takes advantage of the distinction between local and global contexts and outperforms the existing CMs on the performance of automatic metrics (PPL, BLEU, METEOR, NIST, ROUGEL) with significant margins (the best ratios range from 35.49% to 71.61%).

In the next section, we discuss the related works. In Section 3, we present LGCM in detail. In Section 4, we experiment on comparing LGCM with strong baseline CMs. Finally, we make some concluding remarks.

## 2 Related works

We concentrate on the CMs that use transformer-based LMs (see surveys (Tay et al., 2022; de Santana Correia and Colombini, 2022) for transformers and (Bommasani et al., 2021) for LMs). Most CMs use LMs for multi-turn dialogue in open-domain (Wolf et al., 2019; Adiwardana et al., 2020; Roller et al., 2021; Reed et al., 2022; Thoppilan et al., 2022). SOTA CMs were large LMs (LLMs) trained specifically for conversation, such as ChatGPT[2], among other similar models.

Although LLMs can achieve the best practice from time to time, they scale up the Transformer, especially involving concatenating the dialog history into a single sequence. Small models are suitable for the study of CMs first, as the saying goes, it is difficult for a big ship to turn around. Representative CMs are strong baselines based on small LMs such as GPT (Radford et al., 2018) and BERT (Devlin et al., 2019). Among them (Wolf et al., 2019; Zhang et al., 2020; Gu et al., 2021; Wu et al., 2020a; Zhang et al., 2021), TransferTransfo (Wolf et al., 2019) trained especially on the basis of GPT, DialoGPT (Zhang et al., 2020) on GPT2 (Radford et al., 2019), and DialogBERT (Gu et al., 2021) on BERT for dialog response generation.

Hierarchical encoders are a common framework for conversation. HRED was first introduced as two-level RNNs for multi-turn dialogue with a fuse between utterance and context dependencies (Sordoni et al., 2015; Serban et al., 2016, 2017). Most of the attention-based hierarchical models on multi-turn dialogue followed HRED architecture (say (Xing et al., 2018; Tian et al., 2017; Chen et al., 2018; Zhang et al., 2019b,a; Santra et al., 2021), among others). Hierarchical CMs can have different mechanism designs (Zhu et al., 2018; Yang et al., 2019; Li et al., 2020), some of which need an out-of-model mechanism such as learning-to-rank for ranking responses (Cao et al., 2007), for instance, DialogBERT (Gu et al., 2021). There was confusion about the performance between hierarchical versus non-hierarchical (i.e. single level) models. In Lan et al. (2020), hierarchical and non-hierarchical models for open-domain multi-turn dialog generation experienced: hierarchical models were worse than non-hierarchical ones, but hierarchical models with word-level attention were better than non-hierarchical ones. In Santra et al. (2021), it was claimed that hierarchical transformer models with context encoder are effective. Our work proves that hierarchical transformer models are better than non-hierarchical ones without any out-of-model mechanism.

The effectiveness of combining local-global contexts was demonstrated in NLP and CV. It was effective to combine the benefits of using the attention for global context and using the CNN-like or the RNN-like for local context (Yang et al., 2016; Zhang et al., 2019a; Gu et al., 2021; Wu et al., 2020b; Gulati et al., 2020; Wu et al., 2021a; Peng et al., 2022); or using the RNN-like for global context and using the attention for local context (Li et al., 2020). In earlier works, hierarchical transformer encoders use only one token (say `[CLS]`) as the hidden representation of sentence encoding to be fused in the context encoder (say HIBERT (Zhang et al., 2019b), DialogBERT (Gu et al., 2021)). With the dominance of Transformer, it is natural to use Transformer to combine local-global contexts for sequence problems (say (Wu et al., 2021b; Santra et al., 2021; Fang et al., 2022; Hatamizadeh et al., 2023), among others). HIER (Santra et al., 2021) is a strong baseline CM with hierarchical transformer encoders for individual utterances and context respectively, with some limitations compared to our model. In HIER, although contextual embeddings of all utterance tokens are input to the context encoder, the context is a concatenated sequence of utterances in a dialog history. In LGCM, we can separately deal with each utterance in a context and capture full contextualized representations of the local and global contexts by

[2] https://chat.openai.com/

the attention and fuse mechanism.

In essence, the concept of a hierarchical local-global architecture is not a novel one. However, what sets our model apart is our innovative approach to establishing meaningful correlations and connections between local and global contexts. We achieve this by introducing the Inter-attention and Gate mechanisms, which work in tandem to facilitate more coherent and contextually relevant conversations.

## 3 Conversation models

### 3.1 Preliminaries

We write $u = \{u_1, u_2, \cdots, u_T\}$ as a conversation with turn length $T \in \mathbb{N}$, where $\{u_{2k}\}_{k=1}^{\lfloor T/2 \rfloor}$ are utterances from one speaker and $\{u_{2k-1}\}_{k=1}^{\lceil T/2 \rceil}$ are those from the other speaker. We arrange that $u_T$ is the current response and $u_{T-1}$ is the last utterance. We introduce LGCM as an autoregressive generative model by the following equation of conditional distribution for the response $u_T$:

$$P(u_T) = -\sum_{i=1}^{|u_T|} \log P(u_T^i | u_T^{<i}, u_{<T}; f_\theta), \qquad (1)$$

where the conditional probabilities are computed by a neural network that is a (differentiable nonlinear) function $f_\theta$ with parameters $\theta$, which we shall take as a variant of Transformer (Vaswani et al., 2017). The training objective is to maximize the average negative log-likelihood according to Equation 1.

Recall that we distinguish local context for tokens in an utterance at the utterance level and global context for inter-utterances in a dialogue at the dialogue level. We encode local context for each utterance to capture more sensitive information from the neighboring tokens and global context for multiple utterances to capture inter-turn relevance from a dialog history. We obtain contextualized representations of utterances by fusing the local and global contexts.

LGCM is implemented as a local-global encoder-decoder transformer (see Figure 1). We modify the standard transformer encoder as local encoders with PS and global encoder and keep the decoder the same as the standard transformer decoder.

**Embeddings**. Let $e(u_t^i)$ be a single token embedding (i.e. the $i$-th token in the $t$-th utterance), $e(u_t)$ an utterance embedding. We use APE for the token and utterance respectively. Let $p(i)$ be token positional embedding for the $i$-th token that is shared for each utterance and input in the local encoder, and $p_u(t)$ utterance positional embedding for the $t$-th utterance that is input in the global encoder. We use role embedding $r(t)$ for the $t$-th utterance to distinguish whether the speaker is a user or a bot. As usual, we use `[bos]`and `[eos]` as the beginning and end of each utterance to separate between utterances.

We write $\boldsymbol{u}_t^i$ for input representation of token $u_t^i$ as follow:

$$\boldsymbol{u}_t^i = e(u_t^i) + p(i) + r(t). \qquad (2)$$

What follows, we write $\boldsymbol{u}_t$ to denote the utterance embedding $e(u_t) = (\boldsymbol{u}_t^1, \cdots, \boldsymbol{u}_t^{|u_t|})$ for the sake of convenience. We share the input and output embedding matrices as usual done in past practice.

**Local encoder**. We use a standard transformer module as a local encoder of LGCM for each utterance in the local context. The transformer module is stacked layers of the multi-head self-attention followed by the feed-forward with layer normalization in a standard way. For each utterance $u_t$, an utterance representation $\boldsymbol{c}_t = \{\boldsymbol{c}_t^i\}_{i=1}^{|u_t|}$ is produced with the dimension of the value vector of $\boldsymbol{u}_t$, which is a context vector from a self-attention module. The locally contextualized representation $\boldsymbol{c}_t$ essentially summarizes the tokens in $\boldsymbol{u}_t$.

For utterance embeddings $(\boldsymbol{u}_1, \cdots, \boldsymbol{u}_{T-1})$ in the context, the corresponding locally contextualized representations $(\boldsymbol{c}_1, \cdots, \boldsymbol{c}_{T-1})$ is the matrix of context vectors by grouping all the obtained context vectors together as columns.

**Decoder**. We use a standard transformer decoder for LGCM. The decoder is stacked layers of the multi-head self-attention followed by the cross-attention with APE and the feed-forward with layer normalization in a standard way.

### 3.2 Global encoder

We introduce a global encoder of LGCM at the dialogue level. The global encoder comprises the inter-attention and gate mechanism (Figure 1). The hidden representations of the global encoder from the local contexts (Self-Attention) and the global context (Inter-Attention) are fused (via Gate) as the fully contextualized representations of the encoder of LGCM.

For locally contextualized matrix $\boldsymbol{c} = (\boldsymbol{c}_1, \cdots, \boldsymbol{c}_{T-1})$, we write globally contextualized

representation as the matrix $\boldsymbol{C} = (\boldsymbol{C}_1, \cdots, \boldsymbol{C}_{T-1})$ correspondingly. The global representation $\boldsymbol{C}$ models the transformation of global context at the dialogue level from the local representation $\boldsymbol{c}$ at the utterance level as follows:

$$\begin{aligned}\boldsymbol{C} = \text{LayerNorm}(\text{MultiHead}(\\ \text{InterAttention}(\boldsymbol{c}, \boldsymbol{c}, \boldsymbol{c}) + \boldsymbol{c})),\end{aligned} \tag{3}$$

where InterAttention$(Q, K, V)$ is the inter-attention mechanism as described in the following.

**Inter-Attention**. We introduce the inter-attention to extend the attention mechanism to local-global inter-utterance attention by using RPE. The basic idea of InterAttention is that for any turn $t$, $\boldsymbol{c}_t$ attends to all the other $\boldsymbol{c}_s$s in the global context. Our RPE extends the original one (Shaw et al., 2018) from a single sequence in the self-attention to pairwise utterances for the conversation. We use RPE in attention not just for arbitrary pairwise token relations but also arbitrary pairwise utterance relations, which helps capture the structure of conversation in the sense that it refers to the relations between the tokens and utterances in input.

InterAttention$(Q, K, V)$ is defined according to the relation (relative distance) between the $t$-th utterance and the $s$-th utterance as input in the following:

$$\begin{aligned}\boldsymbol{A}_{t,s} &= \frac{1}{\sqrt{d_{out}}}\boldsymbol{c}_t\boldsymbol{W}^Q\left(\boldsymbol{c}_s\boldsymbol{W}^K + \boldsymbol{1}_{|u_s|}\boldsymbol{a}^K_{t,s}\right)^\top,\\ \boldsymbol{C}_t &= \sum_{s=1}^{T-1}\text{Softmax}\left(\boldsymbol{A}_{t,s}\right)\left(\boldsymbol{c}_s\boldsymbol{W}^V\right),\end{aligned} \tag{4}$$

where $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V \in \mathbb{R}^{d_{in}\times d_{out}}$ are matrices to be learned for transforming $\boldsymbol{c}_t$, $\boldsymbol{c}_s$ to their $QKV$-representations, $\boldsymbol{a}^K_{t,s} \in \mathbb{R}^{d_{out}}$ is a learnable vector with the same dimension as $\boldsymbol{c}^j_s\boldsymbol{W}^K$ according to the relative distance between the $t$-th and the $s$-th utterances of the input. Namely, for a query $\boldsymbol{c}^i_t$, the inter-attention computes its globally contextualized representation over all the tokens, $\boldsymbol{c}^j_s$, belonging to their utterances that are locally contextualized representations in the following:

$$\begin{aligned}\boldsymbol{C}^i_t &= \sum_{s=1}^{T-1}\sum_{j=1}^{|u_s|}\alpha^{i,j}_{t,s}(\boldsymbol{c}^j_s\boldsymbol{W}^V),\\ \alpha^{i,j}_{t,s} &= \text{Softmax}(e^{i,j}_{t,s}),\end{aligned} \tag{5}$$

where $\alpha^{i,j}_{t,s}$ is the weight of $\boldsymbol{c}^i_t$ over $\boldsymbol{c}^j_s$. The logit $e^{i,j}_{t,s}$ is computed by the relative distance as follows:

$$e^{i,j}_{t,s} = \frac{1}{\sqrt{d_{out}}}(\boldsymbol{c}^i_t\boldsymbol{W}^Q)(\boldsymbol{c}^j_s\boldsymbol{W}^K + \boldsymbol{a}^K_{t,s})^\top. \tag{6}$$

Notice that we only take the relative distance representation for the key position, $\boldsymbol{a}^K_{t,s}$. As observed in past experiences (Shaw et al., 2018; Huang et al., 2020) and our ablation study, we observe that the key position encoding is key.

In the original RPE, it is assumed that the relative position information is not useful beyond a certain distance and is clipped for the maximum relative position. We take the whole context length as the maximum; that is, we do not need to clip for it. Contrarily, we claim that the relative position information in a dialog history is useful for grounded conversation. The clipped maximum length possible does not allow the conversation to attend over an informative enough context. The global context depends on all the local contexts where information about the relative position representations selected by given attention heads is learnable.

**Gate**. In the global encoder, the Gate follows from the inter-attention for the fusion of Self-Attention in the local context and Inter-Attention in the global context as fully contextualized representations. The fused encoding $\tilde{\boldsymbol{C}}$ is the fuse of the representation $\boldsymbol{c}$ of the local encoders and the one $\boldsymbol{C}$ of the inter-attention by a nonlinear transformation (Sigmoid) for local-global contextualized representation as follows:

$$\begin{aligned}\boldsymbol{H} &= \text{Sigmoid}([\boldsymbol{c}; \boldsymbol{C}]\boldsymbol{W}),\\ \tilde{\boldsymbol{C}} &= (1 - \boldsymbol{H}) \odot \boldsymbol{C} + \boldsymbol{H} \odot \boldsymbol{c},\end{aligned} \tag{7}$$

where $[\boldsymbol{c}; \boldsymbol{C}]$ is the concatenation of $\boldsymbol{c}$ and $\boldsymbol{C}$, $\boldsymbol{W}$ is a learnable linear transformation, $\odot$ indicates element-wise (Hadamard) multiplication. Remember that the fused encoding $\tilde{\boldsymbol{C}}$ outputs to the cross-attention of the decoder.

Finally, a question may be asked whether the structure of LGCM for combining local-global contexts for more informative distribution brings up more computation burden than the Transformer. Most likely, we point out that the computational complexity of LGCM is less than Transformer. Let $L$ be the length of the input sequence and $d$ the dimension of the hidden state. The main computation burden for the single-head transformer encoder layer comes from matrix multiplications of self-attention and feed-forward network (FFN), namely $6Ld^2 + 4L^2d$ for self-attention and $16Ld^2$ for FFN,

respectively. The local encoder of LGCM has the same structure as the Transformer encoder. The difference between them is that the local encoder of LGCM processes each utterance separately, while the Transformer encoder processes the concatenated sequence of utterances. Assume that the input sequence contains $N$ utterances with the same length the computation burden of the self-attention in the local encoder of LGCM is $6Ld^2 + \frac{4L^2d}{N}$, which is more efficient than the Transformer encoder. For comparing the global encoder of LGCM and the Transformer encoder, we first consider the comparison between Inter-Attention and Self-Attention. As shown in Equation 4, the inter-attention adds a deviation about the relative distance to the key, which is negligible compared with matrix multiplication. Thus we consider that the computational complexity of the inter-attention and the self-attention is almost equal. We then consider the comparison between the Gate of LGCM and FFN. Since the computation burden of Sigmoid and element-wise multiplication can be ignored concerning matrix multiplication, the calculation amount of Gate is $4Ld^2$ according to Equation 7, which is more efficient than FFN. To sum up, when the number of layers of both the LGCM encoder and the Transformer encoder is the same, the computational complexity of the LGCM encoder is less. This allows us to scale up the model to a large one.

## 4 Experiments

### 4.1 Setup

**Datasets**. Experiments are conducted on three public-available English multi-turn dialog datasets as follows:

- *PersonaChat* (Zhang et al., 2018): This dataset is randomly paired and asked to get to know each other by chatting according to the given profiles, consisting of 164,356 utterances over 10,981 dialogs.

- *DailyDialog* (Li et al., 2017): This dataset covers a variety of topics in daily life, consisting of 102,979 utterances over 13,118 dialogs.

- *MultiWoz* (Budzianowski et al., 2018): This dataset comprises human-human written conversations in multiple domains and topics, consisting of 115,424 utterances over 8,438 dialogues. Although designed for task-oriented dialogue, the dataset is a good benchmark for open-domain response generation (Gu et al., 2021).

**Comparison models**. We compare LGCM with baseline Transformer (Vaswani et al., 2017), and four strong baseline CMs: TransferTransfo (Wolf et al., 2019), DialoGPT (Zhang et al., 2020), DialogBERT (Gu et al., 2021) and HIER (Santra et al., 2021). Both HIER and LGCM use hierarchical transformer encoders, the comparison between them demonstrates the effectiveness of the global encoder in our model. HIER-CLS (Santra et al., 2021) is a variant of HIER that takes a single token as the embedding for each utterance. We also include HIER-CLS for comparison.

When comparing models, we aim to eliminate the influence of pre-training data and model scale, focusing the comparison on model design. Hence, we re-implement these baseline models to match the scale of LGCM, and then train them on each dataset in a supervised manner. Based on the characteristics of the baseline models, we divide them into two categories. The first group consists of Transformer, HIER, and HIER-CLS, which mainly differ from LGCM in the design of the encoder. To directly reflect the effect of our designs in the LGCM encoder, for models in this group, we use the same input embedding and decoder as LGCM to eliminate the influence of irrelevant factors.[3] The models in the second group, DialoGPT, TransferTransfo, and DialogBERT, all have their special designs. For example, DialoGPT adopted a decoder-only structure, while TranferTransfo employs a multi-task learning paradigm. For these models, we make minimal modifications while retaining model-specific designs of the original models such as input embedding, multi-task learning, and decoding strategy.

**Implementation**. We use the transformers library to implement all the models (Wolf et al., 2020).[4] Transformer consists of 6 encoder layers and 6 decoder layers. All the hierarchical models (DialogBERT, HIER/HIER-CLS, and LGCM) consist of 3 local (or so-called utterance) encoder layers, 3 global (or so-called context) encoder layers, and 6 decoder layers. The decoder-only models (TransferTransfo and DialoGPT) consist of 6 decoder

[3] A subtle distinction is that since the Transformer lacks a hierarchical encoder structure, we add the utterance positional encoding in the input embedding when implementing the Transformer encoder.

[4] https://github.com/huggingface/transformers

| Model | DailyDialog | | | | | MultiWOZ | | | | | PersonaChat | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **PPL** | **BLEU** | **METEOR** | **NIST** | **ROUGEL** | **PPL** | **BLEU** | **METEOR** | **NIST** | **ROUGEL** | **PPL** | **BLEU** | **METEOR** | **NIST** | **ROUGEL** |
| Transformer | 30.03 | 6.86 | 10.61 | 26.48 | 15.61 | 5.01 | 12.95 | 22.05 | 63.62 | 24.05 | 36.66 | 7.65 | 10.52 | 40.95 | 15.77 |
| TransferTransfo | 36.51 | 6.89 | 11.73 | 27.42 | 17.11 | 5.35 | 10.03 | 16.81 | 47.10 | 19.48 | 44.07 | 8.11 | 11.10 | 44.38 | 15.19 |
| DialoGPT | 42.90 | 7.36 | 12.78 | 29.04 | 17.86 | 5.25 | 12.59 | 21.24 | 61.75 | 23.23 | 40.74 | 7.74 | 10.38 | 41.58 | 15.21 |
| DialogBERT | 39.91 | 6.17 | 8.77 | 24.76 | 11.35 | 5.96 | 8.26 | 13.28 | 42.03 | 14.51 | 47.06 | 6.43 | 7.70 | 30.92 | 10.50 |
| HIER | 27.89 | 6.70 | 11.47 | 25.12 | 17.19 | 5.05 | 13.06 | 22.15 | 64.62 | 24.04 | 37.42 | 7.75 | 10.31 | 41.81 | 15.52 |
| HIER-CLS | 30.34 | 6.57 | 11.19 | 25.26 | 16.97 | 5.05 | 12.92 | 21.62 | 65.86 | 23.41 | 39.38 | 7.91 | 10.68 | 43.60 | 15.69 |
| LGCM | **26.48** | **8.36** | **14.08** | **35.56** | **19.17** | **4.99** | **13.26** | **22.79** | **67.66** | **24.24** | **35.87** | **8.41** | **11.79** | **47.07** | **16.73** |

Table 1: Automatic evaluation results on three datasets.

| Model | DailyDialog | | | | | MultiWOZ | | | | | PersonaChat | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **PPL** | **BLEU** | **METEOR** | **NIST** | **ROUGEL** | **PPL** | **BLEU** | **METEOR** | **NIST** | **ROUGEL** | **PPL** | **BLEU** | **METEOR** | **NIST** | **ROUGEL** |
| LGCM | **26.48** | **8.36** | **14.08** | **35.56** | **19.17** | 4.99 | **13.26** | **22.79** | **67.66** | **24.24** | 35.87 | **8.41** | **11.79** | **47.07** | **16.73** |
| -w/o IA | 26.87 | 7.74 | 13.45 | 32.14 | 18.65 | **4.98** | 13.15 | 22.24 | 65.83 | 24.00 | **35.63** | 7.85 | 10.52 | 43.03 | 15.08 |
| -w/o gate | 28.13 | 7.29 | 12.39 | 30.94 | 17.29 | 5.04 | 13.09 | 22.09 | 65.74 | 24.00 | 36.10 | 8.00 | 11.31 | 43.54 | 16.25 |

Table 2: Ablation study results on Inter-Attention and Gate. '- w/o IA' refers to LGCM-w/o Inter-Attention, '- w/o Gate' refers to LGCM-w/o Gate.

layers. The number of attention heads is 8, and the dimension of the hidden state is 512 for all the models. The maximum number of utterances allowed in the context is 7 (Adiwardana et al., 2020; Gu et al., 2021).

The models are optimized by AdamW (Loshchilov and Hutter, 2019). The learning rate is tuned on the validation set, and the model checkpoints that performed best on the validation set are selected for testing. We adopt the sampling strategy for TransferTransfo and DialogBERT during generation as in the original papers. For the other models, we use greedy search.

**Metrics**. The models are evaluated by automatic evaluation metrics as follows:

- *Perplexity* is commonly used in NLP tasks, which measures the ability of a model to predict real samples.
- *BLEU* shows the *N*-gram similarity between the predicted results and the real ones (Papineni et al., 2002). We present BLEU-4 in our experiments.
- *NIST* is an improved version of BLEU that takes into account the amount of information per *N*-gram (Doddington, 2002).
- *METOR* calculates recall in addition to precision and takes into account synonyms (Banerjee and Lavie, 2005).
- *ROUGE-L* measures the similarity between the predicted text and the real one based on the longest common subsequence (Lin, 2004).

## 4.2 Results

### 4.2.1 Evaluation

The automatic evaluation results are shown in Table 1. We see that LGCM performs best on all the metrics with significant margins. The best ratios range from 35.49% to 71.61%, calculated from the table. The results show the effectiveness of LGCM through the fusion of local and global contexts. Therefore, we have positively answered that the distinction between local and global contexts is helpful in conversation.

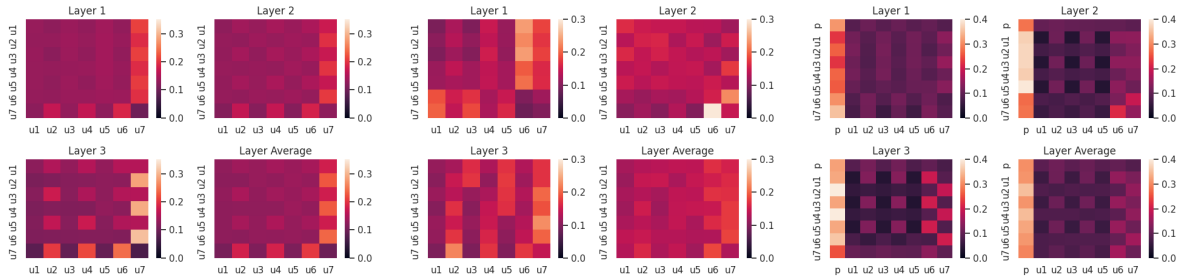
### 4.2.2 Ablation study

To further examine the contributions of the two main designs in the global encoder of LGCM, we conduct ablation studies on Inter-Attention and Gate, respectively. To ensure the computing power of the model, when implementing LGCM-w/o Inter-Attention, we replace Inter-Attention with Self-attention, and when implementing LGCM-w/o Gate, we replace Gate with FFN.

As shown in Table 2, LGCM outperforms LGCM without Inter-Attention on DailyDialog. On the other two datasets, LGCM performs better than LGCM without Inter-Attention except for comparable to PPL. Additionally, removing Gate from LGCM results in a significant performance drop across all the metrics and all the datasets. This study shows that both Inter-Attention and Gate are the proper mechanisms for processing local and global contexts in conversation.

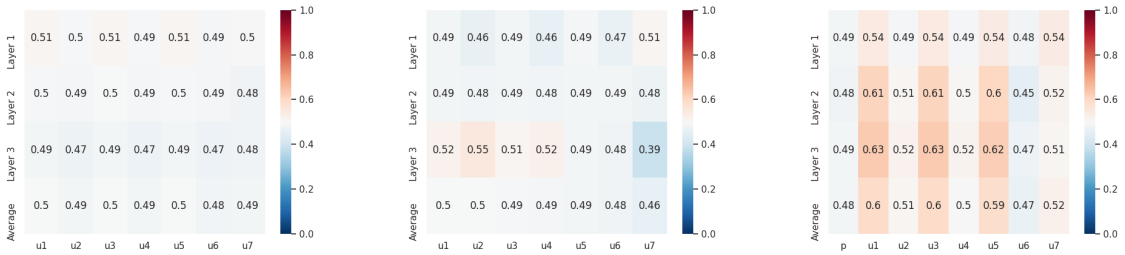
## 4.3 Weight visualization

To figure out how Inter-Attention and Gate help the model understand the contexts, we visualize




(a) DailyDialog (b) MultiWOZ (c) PersonaChat

Figure 2: The attention score visualization of the global encoder on the validation sets. The attention from $u_t$ to $u_s$ is calculated as $a_{t\to s} = \frac{1}{|u_t|}\sum_{i=1}^{|u_t|}\sum_{j=1}^{|u_s|}\alpha_{t,s}^{i,j}$.




(a) DailyDialog (b) MultiWOZ (c) PersonaChat

Figure 3: The gate threshold visualization of the global encoder on validation sets. The values in the heatmap represent the proportion of the global information in the utterance representation, averaged across each token and each hidden dimension.

the attention score and gate threshold in the global encoder of LGCM.

Figure 2 shows the heatmap of the attention weights between utterances. We see that the attention scores between utterances are greatly affected by the utterance's speaker. For example, on the DailyDialog, the last utterance gives greater attention to utterances from partner utterances, especially at deeper layers. Furthermore, historic utterances tend to pay more attention to the latest utterances (the last two turns in our case), which is reasonable since the latest utterances are more relevant to the current dialog topic. In addition, all the historic utterances in PersonaChat have a high attention weight for the persona span, which reflects that the dialogs in the dataset are organized around the given profiles of both participants.

Figure 3 shows the proportion of information from the global representations of utterances. We see that local and global contexts contribute considerably to the representations held among historic utterances and at different layers. This result demonstrates the necessity of using Gate to fuse local and global contexts dynamically. In addition, since Gate has reserved a considerable part of the information for each utterance, an utterance in the attention module usually pays more attention to the context other than itself, thus strengthening the inter-utterance interaction in the entire context.

## 5 Conclusions

Pretrained transformer models are adjusted by concatenating contexts into a single lengthy sequence. It is imperative to explore a variety of methods to encode the context effectively.

We have introduced a local and global conversation model for multi-turn dialogues in open domain. This model harnesses a hierarchical transformer encoder architecture, seamlessly integrating local and global contexts to enhance the efficacy of conversation. We have underscored the significance of distinguishing between the local context for tokens within an utterance at the utterance level and the global context for inter-utterances within a dialogue at the dialogue level. We hope that this study contributes to the comprehension of language models and conversational AI.

## Limitations

LGCM has some limitations. First, it is a small model with limited capability of conversation. We have not experienced scaling it up to a large one and pretraining it on big data. Second, we have not experienced extending it to the cases of multi-modal conversation and multi-task applications. These are areas where LGCM has not been applied, and they can be considered promising directions for future research.

## Acknowledgements

This work was supported by the National Natural Science Foundation of China under grant number 62076009.

## References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*, page 129–136.

Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical variational memory network for dialogue generation. In *Proceedings of the 2018 World Wide Web Conference*, page 1653–1662.

Alana de Santana Correia and Esther Luna Colombini. 2022. Attention, please! a survey of neural attention models in deep learning. *Artificial Intelligence Review*, 55(8):6037–6124.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, page 138–145.

Xiang Fang, Daizong Liu, Pan Zhou, Zichuan Xu, and Ruixuan Li. 2022. Hierarchical local-global transformer for temporal sentence grounding. *arXiv preprint arXiv:2208.14882*.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12911–12919.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Ali Hatamizadeh, Hongxu Yin, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. 2023. Global context vision transformers. In *Proceedings of the 40th International Conference on Machine Learning*, pages 12633–12646.

Zhiheng Huang, Davis Liang, Peng Xu, and Bing Xiang. 2020. Improve transformer models with better relative position embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3327–3335, Online. Association for Computational Linguistics.

Tian Lan, Xian-Ling Mao, Wei Wei, and Heyan Huang. 2020. Which kind is better in open-domain multi-turn dialog, hierarchical or non-hierarchical models? an empirical study. *arXiv preprint arXiv:2008.02964*.

Tianda Li, Jia-Chen Gu, Xiaodan Zhu, Quan Liu, Zhen-Hua Ling, Zhiming Su, and Si Wei. 2020. Dialbert: A hierarchical pre-trained model for conversation disentanglement. *arXiv preprint arXiv:2004.03760*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A manually labelled multi-turn dialogue dataset. In *Proceedings*

*of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Olabiyi Oluwatobi and Erik Mueller. 2020. DLGNet: A transformer-based model for dialogue response generation. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pages 54–62, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Yifan Peng, Siddharth Dalmia, Ian Lane, and Shinji Watanabe. 2022. Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding. In *Proceedings of the 39th International Conference on Machine Learning*, pages 17627–17643.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Lena Reed, Cecilia Li, Angela Ramirez, Liren Wu, and Marilyn Walker. 2022. Jurassic is (almost) all you need: Few-shot meaning-to-text generation for open-domain dialogue. In *Conversational AI for Natural Human-Centric Interaction*.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Chinnadhurai Sankar, Sandeep Subramanian, Chris Pal, Sarath Chandar, and Yoshua Bengio. 2019. Do neural dialog systems use the conversation history effectively? an empirical study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 32–37, Florence, Italy. Association for Computational Linguistics.

Bishal Santra, Potnuru Anusha, and Pawan Goyal. 2021. Hierarchical transformer for task oriented dialog systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5649–5658, Online. Association for Computational Linguistics.

Iulian Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. 2017. Multiresolution recurrent neural networks: An application to dialogue response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, page 3288–3294.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, page 3776–3783.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, New Orleans, Louisiana. Association for Computational Linguistics.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, page 553–562.

Yi Tay, Mostafa Dehghani, Dara Bahri, and Donald Metzler. 2022. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6):1–28.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–236, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, page 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020a. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, and Lei Zhang. 2021a. Cvt: Introducing convolutions to vision transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22–31.

Ting-Wei Wu, Ruolin Su, and Biing-Hwang Juang. 2021b. A Context-Aware Hierarchical BERT Fusion Network for Multi-Turn Dialog Act Detection. In *Proc. Interspeech 2021*, pages 1239–1243.

Zhanghao Wu, Zhijian Liu, Ji Lin, Yujun Lin, and Song Han. 2020b. Lite transformer with long-short range attention. In *International Conference on Learning Representations*.

Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. 2018. Hierarchical recurrent attention network for response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Tianhao Yang, Zheng-Jun Zha, and Hanwang Zhang. 2019. Making history matter: History-advantage sequence training for visual dialog. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2561–2569.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. 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*, pages 1480–1489, San Diego, California. Association for Computational Linguistics.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019a. ReCoSa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730, Florence, Italy. Association for Computational Linguistics.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Xingxing Zhang, Furu Wei, and Ming Zhou. 2019b. HIBERT: Document level pre-training of hierarchical bidirectional transformers for document summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5059–5069, Florence, Italy. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Zhenyu Zhang, Tao Guo, and Meng Chen. 2021. Dialoguebert: A self-supervised learning based dialogue pre-training encoder. In *Proceedings of the 30th ACM International Conference on Information, Knowledge Management*, page 3647–3651.

Chenguang Zhu, Michael Zeng, and Xuedong Huang. 2018. Sdnet: Contextualized attention-based deep network for conversational question answering. *arXiv preprint arXiv:1812.03593*.