Back to the Future: Bidirectional Information Decoupling Network for Multi-turn Dialogue Modeling

Yiyang Li[1,2], **Hai Zhao**[1,2,*] and **Zhuosheng Zhang**[1,2]

[1] Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2] Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University
{eric-lee,zhangzs}@sjtu.edu.cn,zhaohai@cs.sjtu.edu.cn

Abstract

Multi-turn dialogue modeling as a challenging branch of natural language understanding (NLU), aims to build representations for machines to understand human dialogues, which provides a solid foundation for multiple downstream tasks. Recent studies of dialogue modeling commonly employ pre-trained language models (PrLMs) to encode the dialogue history as successive tokens, which is insufficient in capturing the temporal characteristics of dialogues. Therefore, we propose Bidirectional Information Decoupling Network (BiDeN) as a universal dialogue encoder, which explicitly incorporates both the past and future contexts and can be generalized to a wide range of dialogue-related tasks. Experimental results on datasets of different downstream tasks demonstrate the universality and effectiveness of our BiDeN. The official implementation of BiDeN is available at https://github.com/EricLee8/BiDeN.

1 Introduction

Multi-turn dialogue modeling as one of the core tasks in natural language understanding, aims to build representations for machines to understand human dialogues. It is the foundation of solving multiple dialogue-related tasks such as selecting a response (Lowe et al., 2015; Zhang et al., 2018; Cui et al., 2020), answering questions (Sun et al., 2019a; Yang and Choi, 2019; Li et al., 2020a), or making a summarization according to the dialogue history (Gliwa et al., 2019; Chen et al., 2021).

Dialogue contexts possess their intrinsic nature of informal, colloquial expressions, discontinuous semantics, and strong temporal characteristics (Reddy et al., 2019; Yang and Choi, 2019; Chen et al., 2020; Qin et al., 2021a), making them harder for machines to understand compared to plain texts

(Rajpurkar et al., 2016; Cui et al., 2020; Zhang et al., 2021). To tackle the aforementioned obstacles, most of the existing works on dialogue modeling have made efforts from three perspectives. The first group of works adopt a hierarchical encoding strategy by first encoding each utterance in a dialogue separately, then making them interact with each other by an utterance-level interaction module (Zhang et al., 2018; Li and Choi, 2020; Gu et al., 2021). This strategy shows sub-optimal to model multi-turn dialogue owing to the neglect of informative dialogue contexts when encoding individual utterances. The second group of works simply concatenate all the utterances chronologically as a whole (together with response candidates for the response selection task), then encode them using pre-trained language models (PrLMs) (Zhang et al., 2020a; Smith et al., 2020). This encoding pattern has its advantage of leveraging the strong interaction ability of self-attention layer in Transformer (Vaswani et al., 2017) to obtain token-level contextualized embedding, yet ignores utterance-level modeling in dialogue contexts. Sankar et al. (2019) also demonstrate that the simple concatenation is likely to ignore the conversational dynamics across utterances in the dialogue history. The third group of works employ a *pack* and *separate* method by first encoding the whole dialogue context using PrLMs, then separating them to form representations of different granularities (turn-level, utterance-level, etc.) for further interaction (Zhang et al., 2021; Liu et al., 2021a).

Unfortunately, all works mentioned above paid little attention to the temporal characteristics of dialogue texts, which are supposed to be useful and essential for modeling multi-turn dialogues. Different from previous works and to fill the gap of effectively capturing the temporal features in dialogue modeling, we propose a simple but effective Bidirectional Information Decoupling Network (BiDeN), which explicitly incorporates both

Dialogue Context:
#Person1#: *Hi, Della. How long are you going to stay here?*
#Person2#: *Only 4 days. I know that's not long enough, but I have to go to London after the concert here at the weekend.*
#Person1#: *I'm looking forward to your concert very much. Can you tell me where you sing in public for the first time?*
#Person2#: *Hmmm... At my high school concert, my legs shook uncontrollably and I almost fell.*
#Person1#: *I don't believe that. Della, have you been to any clubs in Manchester?*
#Person2#: *No, I haven't. But my boyfriend and I just plan to visit one this evening.*

Downstream tasks:
a). Response selection: (Classification style)
A: **#Person1#:** Great, I can recommend some hotels to you.
B: **#Person1#:** Great, I can recommend some clubs to you.
C: **#Person1#:** What a pity that you are not interested in clubs.
Answer: B
b). Extractive QA: (Retrieval style)
Question: Where the conversation takes place?
Answer: *in Manchester*
c). Dialogue summarization: (Generative style)
Summary: *#Person1# asks Della where she sing in public for the first time and her plans in Manchester.*

Figure 1: An example of different downstream tasks based on dialogue contexts.

the past and future information from the dialogue contexts. Our proposed model can serve as a universal dialogue encoder and be generalized to a wide range of downstream dialogue-related tasks covering classification, retrieval, and generative styles as illustrated in Figure 1.

In detail, we first concatenate all the utterances to form a dialogue context, then encode it with a PrLM. After obtaining the representations output by the PrLM, three additional parameter-independent information decoupling layers are applied to decouple three kinds of information entangled in the dialogue representations: past-to-current, future-to-current, and current-to-current information. Respectively, the past-to-current information guides the modeling of what the current utterance should be like given the past dialogue history, the future-to-current information guides the modeling of what kind of current utterance will lead to the development of the future dialogue, and the current-to-current information guides the modeling of the original semantic meaning resides in the current utterance. After obtaining these representations, we fuse them using a Mixture of Experts (MoE) mechanism (Jacobs et al., 1991) to form the final dialogue history representations.

Let's focus again on Figure 1 and take the response selection task as example. When modeling the three candidate responses, the past-to-current information of the responses and the future-to-

current information of each utterance in the context will detect incoherent temporal features in response *A* and *C*, and coherent feature of response *B*, which help the model to deduce the final answer.

We conduct experiments on three datasets that belong to different types of dialogue-related tasks: Multi-Turn Dialogue Reasoning (MuTual, Cui et al. 2020) for response selection, Molweni (Li et al., 2020a) for extractive question-answering (QA) over multi-turn multi-party dialogues, and DIALOGSUM (Chen et al., 2021) for dialogue summarization. Experimental results on these three datasets show that BiDeN outperforms strong baselines by large margins and achieves new state-of-the-art results.

The contributions of our work are three-fold:

- The proposed model can serve as a universal dialogue encoder and easily be applied to various downstream dialogue-related tasks.
- The proposed model is designed to model the indispensable temporal characteristics of dialogue contexts, which are ignored by previous works. To the best of our knowledge, this is the first paper that introduces the back-and-forth reading strategy (Sun et al., 2019b) to the modeling of temporal characteristics of dialogues.
- Experimental results on three benchmark datasets show that our simple but effective model outperforms strong baselines by large margins, and achieves new state-of-the-art results.

2 Related Works

2.1 Pre-trained Language Models

Our model is implemented based on pre-trained language models (PrLMs), which have achieved remarkable results on many natural language understanding (NLU) tasks and are widely used as a text encoder by many researchers (Wu et al., 2022; Li et al., 2022). Based on self-attention mechanism and Transformer (Vaswani et al., 2017), together with pre-training on large corpora, PrLMs have a strong capability of encoding natural language texts into contextualized representations. To name a few, BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020) and ELECTRA (Clark et al., 2020) are the most prominent ones for NLU; GPT (Radford et al., 2019), T5 (Raffel et al., 2020) and BART (Lewis et al., 2020) are the most representative ones for natural language generation. In our work, we select BERT, ELECTRA, and BART as the encoder backbones of our model.
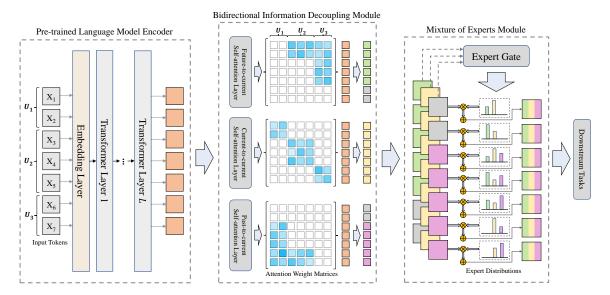
Figure 2: The overview of our model, which consists of three main parts: a pre-trained language model encoder (PrLM encoder), a Bidirectional Information Decoupling Module (BIDM) and a Mixture of Experts (MoE) module. A gray square in the middle part means the representation of this token in this channel is invalid, which will be ignored by the MoE module.

2.2 Multi-turn Dialogue Modeling

There are several previous studies on multi-turn dialogue modeling for different downstream tasks. Li et al. (2021b) propose DialoFlow, which utilizes three novel pre-training objectives to capture the information dynamics across dialogue utterances for response generation. Zhang et al. (2021) design a Pivot-oriented Deep Selection mode (PoDS) to explicitly capture salient utterances and incorporate common sense knowledge for response selection. Liu et al. (2021a) propose a Mask-based Decoupling-Fusing Network (MDFN), which adopts a mask mechanism to explicitly model speaker and utterance information for two-party dialogues. Liu et al. (2021b) propose a Graph Reasoning Network (GRN) to explicitly model the reasoning process on multi-turn dialogue response selection. Different from all these detailed works focusing on specific tasks, in this work, we devote ourselves to a universal dialogue modeling enhancement by effectively capturing the long-term ignored temporal features of dialogue data.

3 Methodology

In this part, we introduce BiDeN and its three modules, whose overview is shown in Figure 2. The left part is a pre-trained language model encoder. Given a sequence of input tokens, the PrLM encoder yields their contextualized representations. The middle part is a Bidirectional Information De-

coupling Module (BIDM), which decouples the entangled representations into three channels for each utterance: future-to-current representations, past-to-current representations and current-to-current representations. The right part is a Mixture of Experts (MoE) module, which calculates an expert distribution to dynamically fuse the three kinds of representations for each token. In the following sections, we will introduce them in detail, respectively.

3.1 Pre-trained Language Model Encoder

Given a set of input tokens $\mathbb{X} = \{w_1, w_2, ..., w_n\}$, we first embed them into a high dimensional embedding space using an embedding look-up table ϕ: $E_T = \phi(\mathbb{X}) = \{e_1, e_2, ..., e_n\} \in \mathcal{R}^{n \times d}$, where d is the hidden size defined by the PrLM. After that, positional embedding E_P and segment embedding E_S will be added to E_T to model the positional and segment information: $E = E_T + E_P + E_S$. E is later fed into the Transformer layers to obtain the contextualized representations H. We first introduce the multi-head self-attention (MHSA) mechanism:

$$
\begin{aligned}
\text{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) &= \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K^T}}{\sqrt{d_k}})\boldsymbol{V} \\
\text{head}_i &= \text{Attn}(E\boldsymbol{W_i^Q}, E\boldsymbol{W_i^K}, E\boldsymbol{W_i^V}) \\
\text{MultiHead}(H) &= [\text{head}_1, \ldots, \text{head}_h]\boldsymbol{W^O}
\end{aligned} \tag{1}
$$

where $\boldsymbol{W}_i^{\boldsymbol{Q}} \in \mathcal{R}^{d \times d_q}$, $\boldsymbol{W}_i^{\boldsymbol{K}} \in \mathcal{R}^{d \times d_k}$, $\boldsymbol{W}_i^{\boldsymbol{V}} \in \mathcal{R}^{d \times d_v}$, $\boldsymbol{W}^{\boldsymbol{O}} \in \mathcal{R}^{hd_v \times d}$ are transformation matrices with trainable weights, h is the number of attention heads, and $[;]$ denotes the concatenation operation. d_q, d_k, d_v are the hidden sizes of the query vector, key vector and value vector, respectively. MHSA is the foundation of Transformer, which is easier to train and can model long distance dependencies. Given the input embeddings E, the Transformer layers $\text{Trans}(E)$ is formulated as follows:

$$
\begin{aligned}
H^0 &= E \in \mathcal{R}^{n \times d} \\
H_{tmp}^i &= \text{LN}(\text{MultiHead}(H^{i-1}) + H^{i-1}) \\
H^i &= \text{LN}(\text{FFN}(H_{tmp}^i) + H_{tmp}^i) \\
\text{FFN}(x) &= \text{ReLU}(x\boldsymbol{W_1} + \boldsymbol{b_1})\boldsymbol{W_2} + \boldsymbol{b_2}
\end{aligned}
\tag{2}
$$

where LN is layer normalization, ReLU is a nonlinear activation function and $\boldsymbol{W_1}$, $\boldsymbol{W_2}$, $\boldsymbol{b_1}$, $\boldsymbol{b_2}$ are trainable linear transformation matrices and bias vectors, respectively.

We denote the stack of L Transformer layers as Trans-L, the final representation output by the PrLM encoder is:

$$
H = \text{Trans-L}(E) \in \mathcal{R}^{n \times d}
\tag{3}
$$

3.2 Bidirectional Information Decoupling

Given the token representations output by the PrLM encoder, the Bidirectional Information Decoupling Module will decouple them into three channels in a back-and-forth way. We first introduce a masked Transformer layer $\text{MTrans}(E, M)$ by modifying the first equation on Eq. (1) to:

$$
\text{Attn}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}(\frac{\boldsymbol{Q}\boldsymbol{K}^{\boldsymbol{T}}}{\sqrt{d_k}} + M)\boldsymbol{V}
\tag{4}
$$

where M is an $n \times n$ attention mask matrix. The function of M is to convert the original fully-connected attention graphs to partially-connected ones, so that each token will be forced to only focus on part of the input sequence. Here we introduce three kinds of attention masks, which guide the decoupling process of the future-to-current channel, current-to-current channel, and past-to-current channel, respectively. Specifically, suppose $I(i)$ means the index of the utterance that the i_{th} token belongs to, the three kinds of masks are obtained by:

$$
\begin{aligned}
M_{f2c}[i,j] &= \begin{cases} 0, & \text{if } I(i) < I(j) \\ -\infty, & \text{otherwise} \end{cases} \\
M_{c2c}[i,j] &= \begin{cases} 0, & \text{if } I(i) = I(j) \\ -\infty, & \text{otherwise} \end{cases} \\
M_{p2c}[i,j] &= \begin{cases} 0, & \text{if } I(i) > I(j) \\ -\infty, & \text{otherwise} \end{cases}
\end{aligned}
\tag{5}
$$

where M_{f2c}, M_{c2c} and M_{p2c} are future-to-current mask, current-to-current mask and past-to-current mask, respectively. After obtaining these masks, three parameter-independent $\text{MTrans-1}(H, M)$ are applied to decouple the original representation H as follows:

$$
\begin{aligned}
H_{f2c} &= \text{MTrans-1}^{\text{f2c}}(H, M_{f2c}) \\
H_{c2c} &= \text{MTrans-1}^{\text{c2c}}(H, M_{c2c}) \\
H_{p2c} &= \text{MTrans-1}^{\text{p2c}}(H, M_{p2c})
\end{aligned}
\tag{6}
$$

Note that there are tokens who has no connections to any tokens in certain channels, e.g. the tokens of the first utterance has no connections to other tokens in past-to-future channel since there are no previous utterances. To handle this case, we simply ignore the invalid representations (gray squares in Figure 2) by adding a fusion mask during the fusion process, which will be introduced in Section 3.3.

After the decoupling process, H_{p2c} contains the information of the influence that the past dialogue history brings about to the current utterance, or in other words, it reflects what the current utterance should be like given the past dialogue history. H_{f2c} contains the information of the influence that the current utterance brings about to future dialogue contexts, or put it another way, it reflects what kind of current utterance will lead to the development of the future dialogue. Finally, H_{c2c} contains the information of the original semantic meaning resides in the current utterance. By explicitly incorporating past and future information into each utterance, our BIDM is equipped with the ability to capture temporal features in dialogue contexts.

3.3 Mixture of Experts Module

We first introduce the Mixture of Experts (MoE) proposed by Jacobs et al. (1991). Specifically, m experts $\{f_i(x)\}_{i=1}^m$ are learned to handle different input cases. Then a gating function $G = \{g_i(x)\}_{i=1}^m$ are applied to determine the importance of each expert dynamically by assigning weights

to them. The final output of MoE is the linear combination of each expert:

$$MoE(x) = \sum_{i=1}^{m} g_i(x) \cdot f_i(x) \qquad (7)$$

In this work, MTrans$^{\text{f2c}}$, MTrans$^{\text{c2c}}$ and MTrans$^{\text{p2c}}$ are treated as three experts. We design the gating function similar as Liu et al. (2021a) that utilizes the original output H to guide the calculation of expert weights. In detail, we first calculate a heuristic matching representation between H and the three outputs of Section 3.2, respectively, then obtain the expert weights G by considering all three matching representations and calculate the final fused representation H_e as follows:

$$
\begin{aligned}
&\text{Heuristic}(X, Y) = [X; Y; X - Y; X \odot Y] \\
&S_f = \text{ReLU}(\text{Heuristic}(H, H_{f2c})\boldsymbol{W_f} + \boldsymbol{b_f}) \\
&S_c = \text{ReLU}(\text{Heuristic}(H, H_{c2c})\boldsymbol{W_c} + \boldsymbol{b_c}) \\
&S_p = \text{ReLU}(\text{Heuristic}(H, H_{p2c})\boldsymbol{W_p} + \boldsymbol{b_p}) \\
&G = \text{Softmax}([S_f; S_c; S_p]\boldsymbol{W_g} + M_g) \in \mathcal{R}^{n \times d \times 3} \\
&H_e = \text{Sum}(\text{Stack}(H_{f2c}; H_{c2c}; H_{p2c}) \odot G)
\end{aligned}
$$
$$(8)$$

Here $H_e \in \mathcal{R}^{n \times d}$, $\odot$ represents element-wise multiplication, $\boldsymbol{W_f}, \boldsymbol{W_c}, \boldsymbol{W_p} \in \mathcal{R}^{4d \times d}$ and $\boldsymbol{b_f}, \boldsymbol{b_c}, \boldsymbol{b_p} \in \mathcal{R}^d$ are trainable transformation matrices and bias vectors, respectively. $\boldsymbol{W_g} \in \mathcal{R}^{3d \times d \times 3}$ is a trainable gating matrix that generates feature-wise expert scores by considering all three kinds of information. M_g is a fusion mask added for ignoring invalid tokens, which is introduced in Section 3.2.

After incorporating future-to-current, past-to-current and current-to-current information, we obtain temporal-aware representation H_e, which can be used for various dialogue-related tasks described in Section 4.

4 Experiments

4.1 Benchmark Datasets

We adopt Multi-Turn Dialogue Reasoning (Mutual, Cui et al. 2020) for response selection, Molweni (Li et al., 2020a) for extractive QA over multi-turn multi-party dialogues, and DIALOGSUM (Chen et al., 2021) for dialogue summarization.

MuTual is proposed to boost the research of the reasoning process in retrieval-based dialogue systems. It consists of 8,860 manually annotated two-party dialogues based on Chinese student English listening comprehension exams. For each dialogue,

four response candidates are provided and only one of them is correct. A *plus* version of this dataset is also annotated by randomly replacing a candidate response with *safe response* (e.g. *I didn't hear you clearly*), in order to test whether a model is able to select a safe response when the other candidates are all inappropriate. This dataset is more challenging than other datasets for response selection since it requires some reasoning to select the correct candidate. This is why we choose it as our benchmark for the response selection task.

Molweni is a dataset for extractive QA over multi-party dialogues. It is derived from the large-scale multi-party dialogue dataset — Ubuntu Chat Corpus (Lowe et al., 2015), whose main theme is technical discussions about problems on the Ubuntu system. In total, it contains 10,000 dialogues annotated with questions and answers. Given a dialogue, several questions will be asked and the answer is guaranteed to be a continuous span in the dialogue context. The reason we choose this dataset as a benchmark for retrieval style task is that we want to test whether our model still holds on multi-party dialogue contexts.

DIALOGSUM is a large-scale real-life dialogue summarization dataset. It contains 13,460 daily conversations collected from different datasets or websites. For each dialogue context, annotators are asked to write a concise summary that conveys the most salient information of the dialogue from an observer's perspective. This dataset is designed to be highly abstractive, which means a generative model should be adopted to generate the summaries.

4.2 Experimental Setups

On the MuTual dataset, ELECTRA is adopted as the PrLM encoder for a fair comparison with previous works. We follow Liu et al. (2021a) to get the dialogue-level representation H_d from H_e. We first obtain the utterance-level representations by applying a max-pooling over the tokens of each utterance, then use a Bidirectional Gated Recurrent Unit (Bi-GRU) to summarize the utterance-level representations into a single dialogue-level vector. For one dialogue history with four candidate responses, we concatenate them to form four dialogue contexts and encode them to obtain $H_D = \{H_d^i\}_{i=1}^{4} \in \mathcal{R}^{d \times 4}$. Given the index of the correct answer i^{target}, we compute the candidate

Model	MuTual			MuTual$^{\text{plus}}$		
	R@1	**R@2**	**MRR**	**R@1**	**R@2**	**MRR**
From Paper (Cui et al., 2020)						
DAM	0.239	0.463	0.575	0.261	0.520	0.645
SMN	0.274	0.524	0.575	0.264	0.524	0.578
BERT	0.657	0.867	0.803	0.514	0.787	0.715
RoBERTa	0.695	0.878	0.824	0.622	0.853	0.782
ELECTRA	0.907	0.975	0.949	0.826	0.947	0.904
+BIDM	*0.916	**0.980**	*0.955	0.830	0.950	0.906
+BiDeN	***0.935**	*0.979	***0.963**	***0.839**	**0.951**	*0.910

Table 1: Results on the development sets of MuTual and MuTual$^{\text{plus}}$. The first four rows are directly taken from the original paper of MuTual. Here $*$ denotes that the result outperforms the baseline model significantly with *p-value* < 0.05 in paired t-test and $**$ denotes < 0.01.

distribution and classification loss by:

$$P_D = \text{Softmax}(\boldsymbol{w_d}^T H_D) \in \mathcal{R}^4$$
$$\mathcal{L}_D = -log(P_D[i^{target}]) \quad (9)$$

where $\boldsymbol{w_d} \in \mathcal{R}^d$ is a trainable linear classifier and $\mathcal{L}_D$ is the cross entropy loss.

On the Molweni dataset, BERT is adopted as the PrLM encoder for a fair comparison with previous works. We simply regard the question text as a special utterance and concatenate it to the end of the dialogue history to form the input sequence. After obtaining H_e, we add two linear classifiers to compute the start and end distributions over all tokens. Given the start and end positions of the answer span $[a_s, a_e]$, cross entropy loss is adopted to train our model:

$$P_{start} = \text{Softmax}(H_e \boldsymbol{w_s}^T) \in \mathcal{R}^n$$
$$P_{end} = \text{Softmax}(H_e \boldsymbol{w_e}^T) \in \mathcal{R}^n$$
$$\mathcal{L}_{SE} = -(\log(P_{start}[a_s]) + \log(P_{end}[a_e]))$$
$$(10)$$

where $\boldsymbol{w_s}$ and $\boldsymbol{w_e} \in \mathcal{R}^d$ are two trainable linear classifiers.

On the DIALOGSUM dataset, BART is chosen as our backbone since it is one of the strongest generative PrLMs. Different from the previous two PrLMs, BART adopts an encoder-decoder architecture where the encoder is in charge of encoding the input texts and the decoder is responsible for generating outputs. Therefore, we add our BIDM after the encoder of BART. Note that BART is pretrained on large corpora using self-supervised text denoising tasks, hence there is a strong coupling on the pre-trained parameter weights between the encoder and decoder. Under this circumstance, simply adding our BIDM after the encoder will destroy the coupling between encoder and decoder, resulting in the decline of model performance. To

tackle this problem, we propose novel a copy-and-reuse way to maintain the parameter-wise coupling between the encoder and decoder. Specifically, instead of using randomly initialized decoupling layers, we reuse the last layer of BART encoder and load the corresponding pre-trained weights to initialize the future-to-current, current-to-current, and past-to-current decoupling layers, respectively. We train this model by an auto-regressive language model loss:

$$\mathcal{L}_G = -\sum_{t=1}^{N} \log p\left(w_t \mid \boldsymbol{\theta}, w_{<t}\right) \quad (11)$$

where $\boldsymbol{\theta}$ is the model parameters, N is the total number of words in the target summary and w_t is the token at time step t. We also conduct experiments on the SAMSum (Gliwa et al., 2019) dataset, and the results are presented in Appendix B.

For hyper-parameter settings and more details about our experiments, please refer to Appendix A.

4.3 Results

In this section, we will briefly introduce the baseline models and evaluation metrics, then present the experimental results on different datasets.

4.3.1 Results on MuTual

Table 1 shows the results on the development sets of MuTual and MuTual$^{\text{plus}}$, respectively. Following Cui et al. (2020), we adopt **R@k** (Recall at K) and **MRR** (Mean Reciprocal Rank) as our evaluation metrics. The baseline models we compare here are: two PrLM-free methods DAM (Zhou et al., 2018) and Sequential Matching Network (SMN, Wu et al. 2017), who encode the context and response separately and match them on different granularities. Three PrLM-based baselines: BERT, RoBERTa (Liu et al., 2019) and ELECTRA.

Model	MuTual / MuTual[plus]		
	R@1	R@2	MRR
GRN	0.915 / 0.841	0.983 / 0.957	0.954 / 0.913
MDFN	0.916 / —	0.984 / —	0.956 / —
DAPO	0.916 / 0.836	**0.988** / 0.955	0.956 / 0.910
CF-DR	0.921 / 0.810	0.985 / 0.946	0.958 / 0.896
BiDeN	**0.930 / 0.845**	0.983 / **0.958**	**0.962 / 0.914**

Table 2: Results on the hidden test sets from the *leaderboard* of MuTual dataset.

Model	EM	F1
BERT$_{DADGraph}$ (Li et al., 2021a)	0.465	0.615
BERT$_{SUP}$ (Li and Zhao, 2021)	0.492	0.640
BERT	0.458	0.602
+BIDM	*0.475	**0.626
+BiDeN	**0.481	**0.632
SUP+BiDeN	**0.503	**0.659

Table 3: Results on Molweni, where * and ** represent the same as in Table 1.

Model	Rouge-1	Rouge-2	Rouge-L
DialoBART	0.533	0.296	0.520
DialSent-PGG	0.547	0.305	**0.535**
BART	0.528	0.289	0.511
+BIDM	0.535	*0.301	*0.523
+BiDeN	**0.548	**0.307	**0.532

Table 4: Results on DIALOGSUM, where * and ** represent the same as in Table 1.

We see from Table 1 that PrLM-free models perform worse than PrLM-based models and different PrLMs have different results, where ELECTRA is the best. Compared with vanilla ELECTRA, simply adding BIDM is able to improve the performance, demonstrating that explicitly incorporating the temporal features has a heavy impact on understanding dialogue contexts. By further equipping BiDeN, we observe giant improvements over ELECTRA by **2.8%** and **1.3%** R@1 on MuTual and MuTual[plus], respectively. Note that the absolute improvements on R@2 are not as high as on R@1. We infer this is because the scores on this metric are already high enough, thus it is harder to achieve very large absolute improvements. However, when it comes to the error rate reduction, BiDeN impressively reduces the error rate from 2.5% to 2.0%, which is a 20% relative reduction.

Table 2 presents the current SOTA models on the leaderboard of MuTual, which is tested on the hidden test set. Graph Reasoning Network (GRN, Liu et al. 2021b) utilizes Graph Convolutional Networks to model the reasoning process. MDFN (Liu et al., 2021a) is introduced in Section 2.2, Dialogue-Adaptive Pre-training Objective (DAPO, Li et al. 2020b) designs a special pre-training objective for dialogue modeling. CF-DR is the previous first place on the leaderboard, but without a publicly available paper. We see from the table that BiDeN achieves new SOTA results on both datasets, especially on MuTual, where we observe a performance gain of **0.9%** R@1 score.

4.3.2 Results on Molweni

Table 3 shows the results on Molweni dataset, where we use **Exactly Match (EM)** and **F1** score as the evaluation metrics. DADGraph (Li et al., 2021a) utilizes the discourse parsing annotations in the Molweni dataset and adopts Graph Neural Networks (GNNs) to explicitly model the discourse structure of dialogues. Compared with them, BiDeN needs no additional discourse labels but per-

forms better. SUP (Li and Zhao, 2021) designs auxiliary self-supervised predictions of speakers and utterances to enhance multi-party dialogue comprehension. We see from the table that our model outperforms vanilla BERT by large margins, which are **2.2%** on EM and **2.5%** on F1, respectively. In addition, SUP can be further enhanced by BiDeN.

4.3.3 Results on DIALOGSUM

Table 4 presents the results on DIALOGSUM. We follow Chen et al. (2021) to adopt **Rouge** (pyrouge) as our evaluation metric, which is widely used in dialogue summarization field (Gliwa et al., 2019; Chen et al., 2021). Rouge-n computes the overlapping ratio of n-grams between the prediction and reference summaries. ROUGE-L computes the longest common subsequence (LCS) between the candidates and references, then calculates the F1 ratio by measuring the recall over references and precision over candidates. Following (Jia et al., 2022), we compute the maximum Rouge score among all references for each sample. Table 4 shows our model again outperforms the strong baseline BART by large margins, with over **2.0%** improvements on all metrics. Besides, compared with the current SOTA models, BiDeN also exhibits its superior capability in summarizing dialogue texts. DialoBART (Feng et al., 2021) utilizes DialoGPT (Zhang et al., 2020b) to annotate keywords, redundant utterances and topic transitions in a dialogue, then explicitly incorporates them into the dialogue texts to train BART. Their work requires annotators to extract additional knowl-

Model	R@1	R@2	MRR
BiDeN	**0.935**	0.979	**0.963**
w/o BIDM	0.912	0.976	0.953
w/o BIDM (same #params)	0.923	**0.981**	0.958
w/o MoE	0.916	0.980	0.955
w/o Bi-GRU	0.927	0.980	0.960

Table 5: Ablation study on development set of MuTual

Model	R@1	R@2	MRR
ELECTRA	0.907	0.975	0.949
+ Bi-LSTM	0.912	0.977	0.952
+ Bi-GRU	0.915	0.978	0.955
+ BiDeN	**0.935**	**0.979**	**0.963**

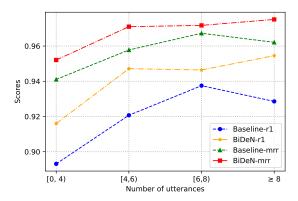Table 6: Results of naive temporal modeling

Figure 3: Model performance v.s. the number of utterances in a dialogue, where the post-fix -r1 represents the **R@1** score and -mrr stands for the **MRR** score.

edge, while our BiDeN still outperforms it on all metrics. DialSent-PGG (Jia et al., 2022) designs a pseudo-paraphrasing process to generate more dialogue-summary pairs from the original dataset, then post-trains the model on the pseudo-summary dataset. After post-training, they fine-tune the summarization model on the original dataset. Compared with their work, which requires an additional post-training process, BiDeN is much simpler and faster to train, yet achieves comparable results.

5 Analysis

In this section, we conduct experiments on MuTual dataset to get an in-depth understanding of BiDeN.

5.1 Ablation Study

To investigate the effectiveness of temporal modeling, we remove BIDM to see how it affects the performance. A sharp performance drop of 2.3% is observed on R@1, demonstrating the necessity and significance of explicit temporal modeling. In order to probe into whether the performance gain comes from the increment of model parameters, we conduct experiments by simply replacing the three kinds of masks defined in Eq. (5) with all-zero masks (fully-connected attention graphs). We see from the table that the increment of parameters does add to the performance. Nevertheless, it is sub-optimal compared with explicitly modeling the temporal features by our BIDM.

We also remove MoE to see whether the dynamic fusion mechanism helps. Specifically, we replace this module with a simple mean pooling over the three decoupled representations. Result shows that MoE makes a huge contribution to the final result. To explore the effect that the task-specific design,

Bi-GRU, brings about to our model, we remove the Bi-GRU and simply average the utterance representations to get the dialogue-level vector. We see from the table that Bi-GRU does have positive effects on the final performance, yet only to a slight extent compared with other modules.

5.2 Naive Temporal Modeling

When it comes to bidirectional temporal modeling, the simplest way is to use Bidirectional Recurrent Neural Networks (Bi-RNNs). To investigate whether BiDeN can be replaced by these naive temporal modeling methods, we conduct experiments by adding Bi-LSTM or Bi-GRU on top of PrLMs instead of BiDeN. We see from Table 6 that utilizing Bi-RNNs can improve the performance slightly, but they are far behind BiDeN. This is because RNNs model the bidirectional information only at token-level, while BiDeN models them by explicitly modeling the utterance boundary with attention masks, which is more consistent with the data characteristics of dialogue texts.

5.3 Influence of Dialogue Length

Intuitively, with longer dialogue contexts comes more complicated temporal features. Based on this point, we analyze the model performance with regard to the number of utterances in a dialogue. As illustrated in Figure 3, the scores first increase from short dialogues to medium-length dialogues. This is because medium-length dialogues contain more information for response matching than short ones. For long dialogues, the baseline model suffers a huge performance drop (see the blue and green lines), while our BiDeN keeps bringing performance improvement, demonstrating a strong ability of it to capture complicated temporal features.

5.4 Visualization of Attentions

To intuitively investigate how BiDeN works, we visualize the attention weights of both current-to-past and current-to-future attentions. Figure 4 (a) shows the current-to-past attention weights. We see that the utterance *My boss told me not to go to work again* tends to focus on *not in a good mood* of the previous utterance, which is a causal discovery. Similarly, the last utterance *I am so sorry that you lost your job* focuses more on *not in a good mood* and *not to go to work*. Figure 4 (b) shows an example of current-to-future attention, which is an incorrect response example taken from MuTual dataset. We see that the current utterance pays great attention on the name *Jane*, which is supposed to be *Joe*. This observation indicates that BiDeN is capable of detecting the logical errors in the future responses that contradict previous utterances. For more visualizations, please refer to Appendix C.
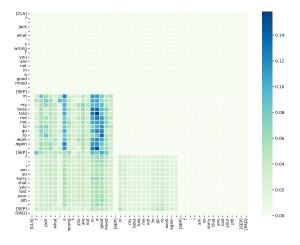
6 Conclusion

In this paper, we propose Bidirectional Information Decoupling Network (BiDeN) to explicitly model the indispensable temporal characteristics of multi-turn dialogues, which have been ignored for a long time by existing works. BiDeN shows simple but effective to serve as a universal dialogue encoder for a wide range of dialogue-related tasks. Experimental results and comprehensive analyses on several benchmark datasets have justified the effectiveness of our model.
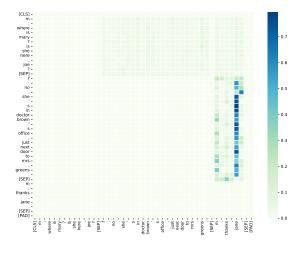
Limitations

Despite the contributions of our work, there are also unavoidable limitations of it.

First, we claim our BiDeN as a universal dialogue encoder which can be used in multiple dialogue-related tasks. In our paper, without the loss of generality, we select three most representative tasks in classification style, retrieval style, and generative style tasks, respectively. However, there are still so many other tasks such as dialogue emotion recognition and dialogue act classification (Qin et al., 2021b), and also so many other large-scale datasets such as Ubuntu, Douban or E-Commerce (Lowe et al., 2015; Zhang et al., 2018; Wu et al., 2017). Due to the lack of computational resources and page limits, our BiDeN is not tested on them. We leave them to the readers who are interested in our model and encourage them to utilize our BiDeN in these tasks.

(a) Current-to-past Attention

(b) Current-to-future Attention

Figure 4: Visualization of attention weights.

Second, the three decoupling layers and the MoE gates add to additional number of parameters (from 348M to 408M), resulting in the increment of computational overheads during training and inference (1.2× slower, 1.2× of GPU memory consumption). However, we argue that the performance gains are worth the additional overheads.

References

Jiaao Chen and Diyi Yang. 2020. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online. Association for Computational Linguistics.

Junfan Chen, Richong Zhang, Yongyi Mao, and Jie Xu. 2020. Neural dialogue state tracking with temporally expressive networks. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1570–1579, Online. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. 2020. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1406–1416, Online. Association for Computational Linguistics.

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.

Xiachong Feng, Xiaocheng Feng, Libo Qin, Bing Qin, and Ting Liu. 2021. Language model as an annotator: Exploring DialoGPT for dialogue summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1479–1491, Online. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 12911–12919. AAAI Press.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive mixtures of local experts. *Neural Comput.*, 3(1):79–87.

Qi Jia, Yizhu Liu, Haifeng Tang, and Kenny Q. Zhu. 2022. Post-training dialogue summarization using pseudo-paraphrasing. *CoRR*, abs/2204.13498.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Changmao Li and Jinho D. Choi. 2020. Transformers to learn hierarchical contexts in multiparty dialogue for span-based question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5709–5714, Online. Association for Computational Linguistics.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020a. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jiaqi Li, Ming Liu, Zihao Zheng, Heng Zhang, Bing Qin, Min-Yen Kan, and Ting Liu. 2021a. Dadgraph: A discourse-aware dialogue graph neural network for multiparty dialogue machine reading comprehension. In *International Joint Conference on Neural Networks, IJCNN 2021, Shenzhen, China, July 18-22, 2021*, pages 1–8. IEEE.

Junlong Li, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2020b. Task-specific objectives of pre-trained language models for dialogue adaptation. *CoRR*, abs/2009.04984.

Yiyang Li, Hongqiu Wu, and Hai Zhao. 2022. Semantic-preserving adversarial code comprehension. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3017–3028, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yiyang Li and Hai Zhao. 2021. Self- and pseudo-self-supervised prediction of speaker and key-utterance for multi-party dialogue reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2053–2063, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021b. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 128–138, Online. Association for Computational Linguistics.

Longxiang Liu, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2021a. Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13406–13414. AAAI Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yongkang Liu, Shi Feng, Daling Wang, Kaisong Song, Feiliang Ren, and Yifei Zhang. 2021b. A graph reasoning network for multi-turn response selection via customized pre-training. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13433–13442. AAAI Press.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Lianhui Qin, Aditya Gupta, Shyam Upadhyay, Luheng He, Yejin Choi, and Manaal Faruqui. 2021a. TIME-DIAL: Temporal commonsense reasoning in dialog. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7066–7076, Online. Association for Computational Linguistics.

Libo Qin, Zhouyang Li, Wanxiang Che, Minheng Ni, and Ting Liu. 2021b. Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13709–13717. AAAI Press.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

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.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2021–2030, Online. Association for Computational Linguistics.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019a. DREAM: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Kai Sun, Dian Yu, Dong Yu, and Claire Cardie. 2019b. Improving machine reading comprehension with general reading strategies. 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 2633–2643, Minneapolis, Minnesota. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Hongqiu Wu, Ruixue Ding, Hai Zhao, Boli Chen, Pengjun Xie, Fei Huang, and Min Zhang. 2022. Forging multiple training objectives for pre-trained language models via meta-learning. *CoRR*.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Zhengzhe Yang and Jinho D. Choi. 2019. FriendsQA: Open-domain question answering on TV show transcripts. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 188–197, Stockholm, Sweden. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020a. 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.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. 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.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Zhuosheng Zhang, Junlong Li, and Hai Zhao. 2021. Multi-turn dialogue reading comprehension with pivot turns and knowledge. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:1161–1173.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127, Melbourne, Australia. Association for Computational Linguistics.

A Hyper-parameter Settings

In this section, we present the detailed hyper-parameter settings of each dataset.

A.1 Hyper-parameters for MuTual

For both MuTual and MuTual$^{\text{plus}}$, we set the maximum input sequence length to 320, where the maximum response length is set to 52 which means the maximum dialogue history length is 268. When truncating the input sequence, we only truncate the dialogue history and leave the response candidates intact. To guarantee the fluency of dialogue history, we truncate them from the front, and at the unit of utterances instead of tokens. The learning rate, training epochs, and batch size are set to 6e-6, 3, and 2, respectively. We use AdamW as our training optimizer and a linear scheduler to schedule the learning rate. The learning rate is first linearly warmed up from 0 to 6e-6 at the first 1% steps then decreased linearly to 0 until the end of training.

A.2 Hyper-parameters for Molweni

For the Molweni dataset, the maximum input sequence length is set to 384, where the maximum question length is 32. Similar to the MuTual dataset, we only truncate the dialogue history and leave the question sentence intact. The learning rate, training epochs, and batch size are set to 7e-5, 5, and 16, respectively. As for the optimizer and scheduler, they are the same as the ones on MuTual dataset.

A.3 Hyper-parameters for DIALOGSUM

For the DIALOGSUM dataset, the maximum input sequence length and maximum summary length are set to 512 and 100, respectively. The learning rate, training epochs, and batch size are set to 2e-5, 15, and 12, respectively. During inference, we use beam search to generate summaries, and set the beam size to 4.

B Results on SAMSum Dataset

For the dialogue summarization task, we also conduct experiments on the SAMSum (Gliwa et al., 2019) dataset. SAMSum is a dialogue summarization dataset that contains 16,369 dialogues in the form of online chatting messages. Compared with DIALOGSUM, which is taken from real-life person-to-person conversations, this dataset contains dialogues that are more informal and colloquial. However, the summaries in this dataset are

Model	Rouge-1	Rouge-2	Rouge-L
Multi-View BART	0.534	0.280	0.499
DialSent-PGG	0.535	0.289	0.502
DialoBART	0.537	0.288	0.508
ConDigSum	**0.542**	0.289	**0.509**
BART	0.526	0.271	0.492
+BIDM	*0.531	*0.278	*0.498
+BiDeN	**0.540	****0.291**	**0.506

Table 7: Results on SAMSum, where * and ** represent the same as in Table 1.

less abstractive than DIALOGSUM (Chen et al., 2021).

Results on SAMSum are tabulated in Table 7, where we can see that BiDeN consistently outperforms the strong baseline BART by large margins. We also compare BiDeN with different models that are also built on BART. Multi-View BART (Chen and Yang, 2020) incorporates different information like topic and stage of dialogues to generate summaries using a multi-view decoder. ConDigSum is the current SOTA model on the SAMSum dataset, which designs two contrastive auxiliary tasks: Coherence Detection and Sub-summary Generation to implicitly model the topic information of dialogues. This model is trained with an alternating updating strategy, which is approximately three times slower than our BiDeN during training since it requires three backward calculations in a single batch. DialoBART and DialSent-PGG are introduced in Section 4.3.3. Table 7 shows that BiDeN achieves comparable results to ConDigSum and outperforms all other models. It is worth noting that all of the previous models require additional dialogue annotators or training stages, while our BiDeN is annotator-free, plug-and-play, and easy to use.

Note that the original results of Multi-View and ConDigSum are obtained by the files2rouge package based on the official ROUGE-1.5.5.pl Perl script, while DialoBART and DialSent-PGG adopt py-rouge. To make fair comparisons, we download the output predictions of Multi-View and ConDig-Sum, then run the py-rouge script to get the corresponding results, which are the ones presented in Table 7.

For the SAMSum dataset, we set the maximum dialogue history length to 800, and the maximum summary length to 100. The learning rate, training epochs, and batch size are set to 2e-5, 5, and 4, respectively. We also adopt beam search during inference, where the beam size is also set to 4.
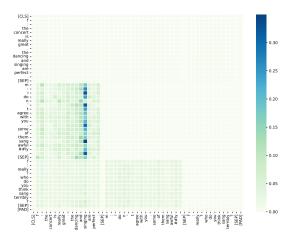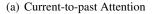
C More Visualizations

We present more examples of the three kinds of attentions: current-to-past attention, current-to-future attention, and current-to-current attention, for readers to further explore how BiDeN works.
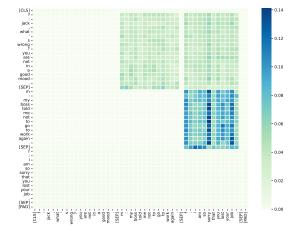
Figure 5 (a) illustrates a conversation about a concert, where the female thinks the dancing and singing are perfect but the male disagrees. We can see from the attention weights that when modeling the second utterance, BiDeN focuses mostly on *dancing and singing*, especially on *singing*, which is consistent with its semantic meaning that some singers sang awfully. In other words, BiDeN is capable of extracting the key information of previous utterances when modeling the current utterance.

Figure 5 (b) is another example of Current-to-future attention, where the male is unhappy because he lost his job and the female feels sorry about that. It can be observed that when modeling the second utterance, BiDeN attends more on *sorry* and *you lost your job*. This observation demonstrates that BiDeN is able to locate the key information in the future utterances to model what kind of current utterance will lead to the development of the future dialogue.
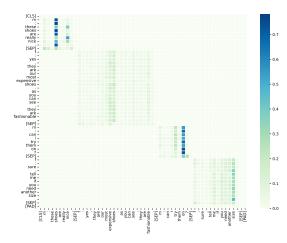
Figure 5 (c) shows an example of current-to-current attention, which is the self-attention within each utterance. Let's focus on each utterance. The first utterance mainly attends to *shoes* and *nice*, which are two keywords that best reflect the semantic meaning of this utterance. Similar observations can be seen in the rest three utterances, where the most prominent words are *expensive shoes* and *fashionable*, *try on*, and *you need another size*, respectively. This observation indicates that BiDeN can model the most salient and concise semantic meaning in each utterance.

(a) Current-to-past Attention

(b) Current-to-future Attention

(c) Current-to-current Attention

Figure 5: More visualization results.