Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances

Zekang Li[12], **Jinchao Zhang**[3], **Zhengcong Fei**[12], **Yang Feng**[12]*, **Jie Zhou**[3]
[1] Key Laboratory of Intelligent Information Processing
Institute of Computing Technology, Chinese Academy of Sciences (ICT/CAS)
[2] University of Chinese Academy of Sciences
[3] Pattern Recognition Center, WeChat AI, Tencent Inc, China
{lizekang19g,feizhengcong,fengyang}@ict.ac.cn
{dayerzhang,withtomzhou}@tencent.com

Abstract

Nowadays, open-domain dialogue models can generate acceptable responses according to the historical context based on the large-scale pre-trained language models. However, they generally concatenate the dialogue history directly as the model input to predict the response, which we named as the *flat pattern* and ignores the dynamic information flow across dialogue utterances. In this work, we propose the **DialoFlow** model, in which we introduce a dynamic flow mechanism to model the context flow, and design three training objectives to capture the information dynamics across dialogue utterances by addressing the semantic influence brought about by each utterance in large-scale pre-training. Experiments on the multi-reference Reddit Dataset and DailyDialog Dataset demonstrate that our DialoFlow significantly outperforms the DialoGPT on the dialogue generation task. Besides, we propose the **Flow score**, an effective automatic metric for evaluating interactive human-bot conversation quality based on the pre-trained DialoFlow, which presents high chatbot-level correlation ($r = 0.9$) with human ratings among 11 chatbots. Code and pre-trained models will be public. [1]

1 Introduction

Recent intelligent open-domain chatbots (Adiwardana et al., 2020; Bao et al., 2020; Smith et al., 2020) have made substantial progress thanks to the rapid development of the large-scale pre-training approaches (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020) and the large amount of conversational data (Dinan et al., 2019; Baumgartner et al., 2020; Smith et al., 2020). However, effectively modeling the dialogue history in large-scale dialogue pre-training is still challenging.

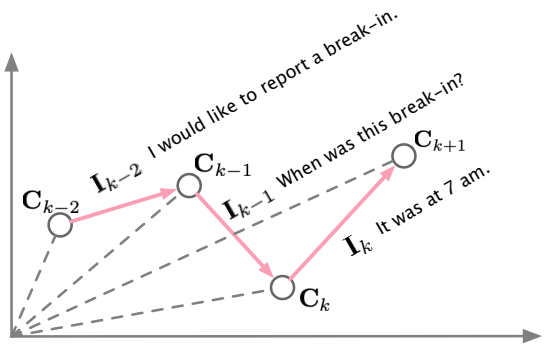

Figure 1: Illustration of the dynamic information flow in the semantic space. $\mathbf{C}_k$ (gray dotted line) denotes the dense representation of dialogue history which we named as context. $\mathbf{I}_k$ (pink solid line) denotes the semantic influence brought about by the k-th utterance, which is the difference between $\mathbf{C}_k$ and $\mathbf{C}_{k+1}$.

Most of the previous work on dialogue history modeling mainly fall into two groups. One group of works generally concatenate the dialogue history as the model input and predict the response (Zhang et al., 2020; Smith et al., 2020; Bao et al., 2020), named as *flat pattern*, which is commonly adopted in the large-scale pre-training. However, Sankar et al. (2019) demonstrate that flat concatenation is likely to ignore the conversational dynamics across utterances in the dialogue history. Another group of works employ *hierarchical modeling* to encode the dialogue history (Serban et al., 2016b; Shan et al., 2020; Gu et al., 2020), in which the utterances are separately encoded and then fed into an utterance-level encoder. These approaches lack the history information when encoding each individual utterance, while the history information is essential for understanding dialogue utterances. Thus, all the aforementioned methods are deficient in modeling the dynamic information in the dialogue history.

In this work, inspired by the human cognitive

*Joint work with Pattern Recognition Center, WeChat AI, Tencent Inc. Yang Feng is the corresponding author. Work was done when Zekang Li and Zhengcong Fei were intern at WeChat AI.

[1] https://github.com/ictnlp/DialoFlow

process that humans always consider the goal or influence of the next response before they continue the conversation (Brown-Schmidt and Konopka, 2015), we propose the DialoFlow to model the dynamic information flow in the dialogue history by addressing the semantic influence brought about by each utterance. As shown in Figure 1, we define the dense representation of the dialogue history at different utterances as the *context*s (gray dot line) and the context transformation as the *semantic influence* brought by each utterance. In particular, our DialoFlow constructs the process of the utterance-level history context flow. Correspondingly, the semantic influence of each utterance can be measured by the difference between two adjacent contexts, which will be further used to guide the current response generation.

Practically, we first employ a transformer to encode the whole conversation to get the dense context representation. Then we design a unidirectional Flow module to capture the context flow on the utterance level, and design three training objectives to model the context flow and measure the semantic influence brought about by each utterance: 1) *Context Flow Modeling*, which aims to capture the context flow schema. 2) *Semantic Influence Modeling*, which targets to measure the predicted semantic influence. 3) *Response Generation Modeling*, which is to generate the response under the guidance of the predicted semantic influence. Furthermore, to demonstrate the effect of modeling dynamic information flow in the dialogue understanding, we propose the **Flow score** based on the DialoFlow, an automatic reference-free evaluation metric for interactive dialogue evaluation by measuring the semantic influence perplexity.

We pre-train the proposed DialoFlow on the large-scale Reddit comments and conduct experiments on dialogue generation and interactive dialogue quality evaluation. For dialogue generation, DialoFlow achieves significant improvements on the Reddit multi-reference dataset and the DailyDialog dataset compared to the baseline DialoGPT (Zhang et al., 2020). For interactive dialogue quality evaluation, our proposed Flow score obtains an impressively high chatbot-level correlation ($r = 0.9$) with human ratings on 2200 human-bot dialogues from 11 chatbots.

Our contributions are summarized as follows:

- We propose the **DialoFlow**, a new paradigm to construct the dynamic information flow in the dialogue history by addressing the semantic influence brought about by each utterance. Besides, we design an automatic reference-free evaluation metric **Flow score** based on the pre-trained DialoFlow for interactive dialogue quality evaluation.
- The experimental results illustrate that DialoFlow achieves significant improvements on dialogue generation compared to the DialoGPT, and Flow score shows impressively high chatbot-level correlation ($r = 0.9$) with human ratings.

2 Method

The proposed DialoFlow models the dynamic information flow in the whole dialogue history by addressing the semantic influence brought about by each utterance in sequence.

2.1 Model Overview

Before introducing the DialoFlow in detail, we first define some terms. Formally, let $\mathcal{D} = \{u_1, u_2, ..., u_N\}$ denotes a whole dialogue. And for each utterance $u_k = \{u_k^1, u_k^2, ..., u_k^T\}$ where u_k^t denotes the t-th word in the k-th utterance. We further denote $u_{<k} = \{u_1, u_2, ..., u_{k-1}\}$ as the dialogue history at the k-th utterance. Besides, the dense representation of the dialogue history $u_{<k}$ at the k-th utterance is represented as the **context** $\mathbf{C}_k$. And the difference between the new context $\mathbf{C}_{k+1}$ at the (k+1)-th utterance and the previous contexts $\mathbf{C}_k$ at the k-th utterance can be defined as the **semantic influence** $\mathbf{I}_k$ of the k-th utterance, which can be formulated as:

$$\mathbf{I}_k = \mathbf{C}_{k+1} - \mathbf{C}_k. \tag{1}$$

In our method, DialoFlow first encodes the dialogue history and predicts the future context $\mathbf{C}'_{k+1}$ according to all the previous history context $\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_k$. Then at the response generation stage, the model acquires the predicted target semantic influence $\mathbf{I}'_k$, and generate the target response u_k auto-regressively considering both the predicted semantic influence and the historical subsentences. Specifically, as shown in Figure 2, DialoFlow models the context flow by designing a unidirectional Flow module upon the transformer, and we introduce three multi-task training objectives to supervise the context flow, semantic influence, and

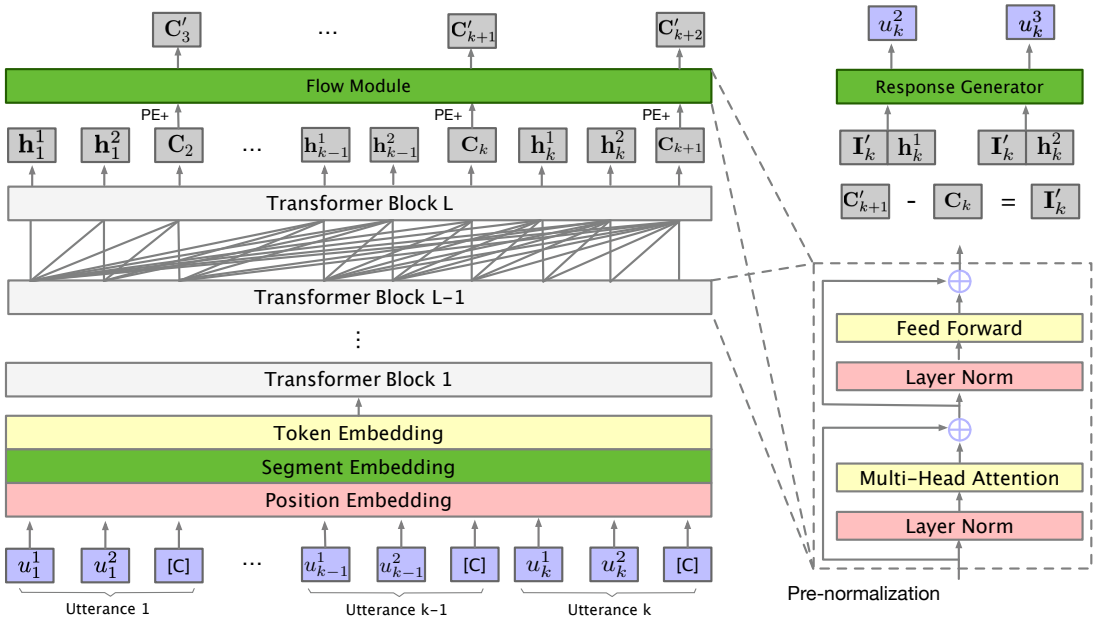

Figure 2: Overview of our DialoFlow. We present the detail model architecture, and the self-attention visualization in DialoFlow. "[C]" is a special token placed at the end of each utterance to model the dense representation of dialogue history $\mathbf{C}_k$ named as the context. The future context $\mathbf{C}'_{k+1}$ can be predicted by context history $[\mathbf{C}_1, \ldots, \mathbf{C}_k]$. For the simplicity, we only plot two tokens for each utterance.

response generation, which are referred to as *context flow modeling*, *semantic influence modeling*, and *response generation modeling*, respectively.

2.2 Model Architecture

Figure 2 demonstrates the infrastructure of DialoFlow, which consists of the input embeddings, transformer blocks, a uni-directional Flow module, and a response generator.

Input Embedding. DialoFlow takes the sum of token embedding, segment embedding, and position embedding as the model input. In particular, we insert a special token "[C]" at the end of each utterance, which is used to capture the overall dense representation of the dialogue history. To enhance the modeling of different speakers, we utilize segment embedding containing two types: "[Speaker1]" and "[Speaker2]".

Transformer Block. A transformer block consists of the following key components: layer normalization, multi-head attention, and feed-forward layers. We employ the pre-normalization used in GPT-2 (Radford et al., 2019) instead of the post-normalization used in BERT (Devlin et al., 2019), as (Shoeybi et al., 2019) show that the post-normalization leads to performance degradation when the model size increases while pre-normalization enables stable large-scale training. DialoFlow keeps the uni-directional dialogue encoding and enables training on the dialogue level rather than on the context-response setting. We can obtain the history context at the k-th utterance encoded by the transformer blocks:

$$\mathbf{C}_k = \text{Transformer}(u_{<k}), \tag{2}$$

where $\mathbf{C}_k$ is the hidden states at the position of special token "[C]". And the hidden states at the position of each token u_k^t in the input sequence are denoted as $\mathbf{h}_k^t$.

Flow Module. To capture the dynamic information flow across the dialogue utterances, we design a Flow module to model the context changing scheme. The architecture of the Flow module is the same with one layer of transformer block. The Flow module takes all the previous context $\{\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_k\}$ as input and predicts the context at the (k+1)-th utterance $\mathbf{C}'_{k+1}$:

$$\mathbf{C}'_{k+1} = \text{Flow}(\mathbf{C}_1, \mathbf{C}_2, ..., \mathbf{C}_k). \tag{3}$$

The predicted semantic influence brought about by the k-th utterance can be computed as:

$$\mathbf{I}'_k = \mathbf{C}'_{k+1} - \mathbf{C}_k. \tag{4}$$

Response Generator. DialoFlow generates the utterance u_k with the guidance of the predicted semantic influence $\mathbf{I}'_k$. The response generator contains a feed-forward layer and a softmax layer to

convert the hidden states to tokens. When generating the t-th word, the response generator takes the predicted semantic influence $\mathbf{I}'_k$ and the hidden states $\mathbf{h}_k^{t-1}$ as input, and outputs the probability distribution of the t-th word:

$$p(u_k^t|\mathbf{I}'_k, u_{<k}, u_k^{<t}) = \text{softmax}(W_1[\mathbf{I}'_k; \mathbf{h}_k^{t-1}] + b_1) \in \mathbb{R}^{|V|}, \quad (5)$$

where $|V|$ refers to the vocabulary size, W_1 and b_1 are learnable parameters.

2.3 Training Objectives

Different from traditional training approaches with context-response pair, DialoFlow is trained with the whole dialogue containing N utterances. Correspondingly, we design three training tasks to optimize the model: 1) Context Flow Modeling, 2) Semantic Influence Modeling, and 3) Response Generation Modeling.

Context Flow Modeling. To capture the dynamic context flow, DialoFlow predicts the context at the k-th utterance $\mathbf{C}'_k$ based on the previous context sequence $\{\mathbf{C}_1, ..., \mathbf{C}_{k-1}\}$. We minimize the L2 distance between the predicted context $\mathbf{C}'_k$ and the real context $\mathbf{C}_k$:

$$\mathcal{L}_{CFM} = \sum_{k=1}^{N} ||\mathbf{C}_k - \mathbf{C}'_k||_2^2. \quad (6)$$

Semantic Influence Modeling. To force the effectively modeling of semantic influence brought about by the n-th utterance at the context $\mathbf{C}_{n-1}$, we design a bag-of-words loss using the predicted semantic influence $\mathbf{I}'_n$:

$$\mathcal{L}_{SIM} = -\sum_{k=1}^{N}\sum_{t=1}^{T} \log p(u_k^t|\mathbf{I}'_k) = -\sum_{k=1}^{N}\sum_{t=1}^{T} \log f_{u_k^t}, \quad (7)$$

where $f_{u_k^t}$ denotes the estimated probability of the t-th word u_k^t in the utterance u_k. The function f is used to predict the words in the utterance u_k in a non-autoregressive way:

$$f = \text{softmax}(W_2\mathbf{I}'_k + b_2) \in \mathbb{R}^{|V|}, \quad (8)$$

where $|V|$ refers to the vocabulary size, W_2 and b_2 are learnable parameters.

Response Generation Modeling. The predicted semantic influence $\mathbf{I}'_k$ can also be regarded as a semantic expectation of the k-th utterance. We incorporate the predicted semantic influence $\mathbf{I}'_k$ into the response generation stage to guide the generation. The response generation objective is as follows:

$$\mathcal{L}_{RGM} = -\sum_{k=1}^{N} \log p(u_k|\mathbf{I}'_k, u_{<k}) = -\sum_{k=1}^{N}\sum_{t=1}^{T} \log p(u_k^t|\mathbf{I}'_k, u_{<k}, u_k^{<t}). \quad (9)$$

The overall training objective of DialoFlow can be computed as follows:

$$\mathcal{L} = \mathcal{L}_{CFM} + \mathcal{L}_{SIM} + \mathcal{L}_{RGM}. \quad (10)$$

2.4 Flow Score

By optimizing with the aforementioned three training objectives, DialoFlow can capture the dynamic information flow across the dialogue history. As the DialoFlow is trained on human-human dialogues, the context flow scheme can be regarded as the general expectation of the dialogue development. Therefore, the closer gap between the semantic influence brought by the chatbot's utterance and the expectation means the more human-likeness.

Based on the consideration, we propose an automatic reference-free metric **Flow score** for interactive dialogue evaluation based on DialoFlow. In the human-bot conversation, when the bot generates a new utterance u_k, we measure the similarity between the predicted semantic influence $\mathbf{I}'_k$ and the real semantic influence $\mathbf{I}_k$ brought about by the utterance u_k, which can be considered as the probability of the human-likeness of the utterance. To compute the similarity between the semantic influences, we measure both the cosine similarity and the length similarity:

$$s_k = \cos(\langle \mathbf{I}'_k, \mathbf{I}_k \rangle) \cdot length(\mathbf{I}'_k, \mathbf{I}_k) = \frac{\mathbf{I}'_k \cdot \mathbf{I}_k}{||\mathbf{I}'_k||\ ||\mathbf{I}_k||} \cdot \frac{\min(||\mathbf{I}'_k||, ||\mathbf{I}_k||)}{\max(||\mathbf{I}'_k||, ||\mathbf{I}_k||)}. \quad (11)$$

Note that we introduce the length similarity to consider the influence of length difference on semantic similarity. For the overall quality of the chatbot in the dialogue, we design a metric, which can be regarded as the dialogue-level perplexity:

$$Flow\ score = 2^{-\frac{1}{M}\sum_k^M \log(\frac{s_k+1}{2})}, \quad (12)$$

where M denotes the turn numbers of the chatbot utterances and $\frac{s_k+1}{2}$ is to scale the similarity value to $[0, 1]$. A lower Flow score corresponds to better dialogue quality.

3 Experiments

3.1 Dataset

For model pre-training, we use the Reddit comments, which are collected by a third party and made publicly available on pushshift.io (Baumgartner et al., 2020). We clean the data following the pipeline used in the DialoGPT.[2]

For response generation, we employ the multi-reference Reddit Test Dataset (Zhang et al., 2020) which contains 6k examples with multiple references. We evaluate our pre-trained DialoFlow model on this dataset. The average length of the dialogue history in this dataset is 1.47. To further explore the dynamic information flow in the long dialogue history situation, we choose another popular open-domain dialogue dataset – DailyDialog Dataset (Li et al., 2017), in which the average dialogue history length is about 4.66. DialoFlow is fine-tuned on the DailyDialog training set and evaluated on the DailyDialog multi-reference test set (Gupta et al., 2019).

For interactive dialogue quality evaluation, we employ the collected data from the Interactive Evaluation of Dialog Track @ The Ninth Dialog System Technology Challenge (DSTC9) (Gunasekara et al., 2021), which contains 2200 human-bot conversations from 11 chatbots. For each conversation, there are 3 human ratings on the overall quality (0-5). We calculate the correlation between the results of our proposed metric and the human ratings on the chatbot level. Human-human conversations are always regarded to be better than human-bot conversations. Therefore, we randomly sample 200 human-human dialogues from the BST (Smith et al., 2020) dataset to see the metric's performance on the real human-human conversations.

3.2 Experimental Setting

Pre-training Details. DialoFlow is pre-trained based on the pre-trained GPT-2 (Radford et al., 2019), since Zhang et al. (2020) show that DialoGPT trained from the pre-trained GPT-2 is much better than from scratch. There are three different model sizes: DialoFlow-base, DialoFlow-medium, and DialoFlow-large, which are trained from the pre-trained GPT2-base, GPT2-medium, GPT2-large, respectively. We used AdamW optimizer (Loshchilov and Hutter, 2019) with 0.01 weight decay and linear learning rate scheduler with 12000 warm-up steps. The learning rate is 2e-4 for the base and medium version and 1e-4 for the large version. We use the batch size of 1024 for all model sizes. We trained the base and medium models for up to 4 epochs and trained the large model for 2 epochs. It costs about two months on 8 Nvidia V100 GPUs to train the large model.

Decoding Details. On the 6K Reddit multi-reference dataset, we use beam search (with beam width 10) on the DialoFlow-medium model and the DialoFlow-large model. We employ greedy search on the DialoFlow-base model, which keeps the same with (Zhang et al., 2020). On the DailyDialog dataset, we fine-tune the pre-trained DialoFlow and DialoGPT, select the checkpoint based on the validation loss, and then use beam search (with beam width 5) for decoding.

3.3 Baseline

For response generation, we compare our proposed DialoFlow with DialoGPT, a popular dialogue generation model pre-trained on the Reddit Comments. We choose the version trained from pre-trained OpenAI GPT-2 for comparison.

For interactive dialogue evaluation, we compare our metric with the following metrics: 1) **FED score** (Mehri and Eskénazi, 2020) is an automatic evaluation metric which uses DialoGPT-large, without any fine-tuning or supervision. FED takes the DialoGPT-large as the user and calculates the likelihood of follow-up utterances based on several pre-set usual human utterances. FED works under the pre-set common human utterances, which can reveal the dialogue quality. 2) **Perplexity** is used to measure the coherence of an utterance under the dialogue context. We employ DialoGPT-large to measure the perplexity for each utterance of the chatbot. We average the perplexity of all utterances in the whole dialogue as the baseline metric.

3.4 Evaluation Metrics

For dialogue response generation, we perform automatic evaluation using common reference-based metrics: BLEU (Papineni et al., 2002), METEOR (Lavie and Agarwal, 2007), and NIST (Lin and Och, 2004). NIST is a variant of BLEU that weights n-gram matches by their information gain, i.e., it indirectly penalizes uninformative n-grams

[2] https://github.com/microsoft/DialoGPT

Method	NIST-2	NIST-4	BLEU-2	BLEU-4	METEOR	Entropy	Avg Len
Multi-reference Reddit Dataset							
DialoGPT (B, greedy)	2.39	2.41	10.54%	1.55%	7.53%	**10.77**	12.82
DialoFlow (B, greedy)	2.88	2.93	15.34%	3.97%	9.52%	9.27	**15.43**
DialoGPT (M, beam)	3.40	3.50	**21.76%**	**7.92%**	10.74%	10.48	11.34
DialoFlow (M, beam)	3.89	3.99	20.98%	7.36%	11.46%	10.42	13.37
DialoGPT (L, beam)	2.90	2.98	21.08%	7.57%	10.11%	10.06	10.68
DialoFlow (L, beam)	**3.90**	**4.01**	21.20%	7.42%	**11.48%**	10.42	13.38
Human	3.41	3.50	17.90%	7.48%	10.64%	10.99	13.10
Multi-reference DailyDialog Dataset							
DialoGPT (B, beam)	2.28	2.78	18.83%	6.63%	15.5%	9.80	**18.82**
DialoFlow (B, beam)	3.65	3.84	26.47%	10.12%	16.1%	9.62	12.00
DialoGPT (M, beam)	3.47	3.65	25.39%	9.99%	15.9%	9.64	12.88
DialoFlow (M, beam)	3.80	4.02	27.63%	11.33%	16.7%	9.83	12.06
DialoGPT (L, beam)	3.30	3.46	23.69%	9.20%	15.7%	9.78	13.24
DialoFlow (L, beam)	**3.86**	**4.08**	**28.02%**	**11.57%**	**17.0%**	**9.87**	12.08
Ablation Study on Multi-reference Reddit Dataset							
DialoFlow (M, beam)	3.89	3.99	20.98%	7.36%	11.46%	10.42	13.37
w/o SIM	3.85	3.96	21.36%	7.71%	11.26%	10.43	12.70
w/o SIM & CFM	3.79	3.89	21.33%	7.65%	11.25%	10.33	12.55

Table 1: The evaluation on 6K Reddit multi-reference dataset and on DailyDialog dataset. For 6K Reddit multi-reference dataset, as the DialoGPT do not release the decoding code, we directly quote the results from (Zhang et al., 2020). Note that "B", "M", "L" denotes base, medium, large respectively.

Metric	DialoFlow	DialoGPT	Tie
Relevance	43.7%	28.8%	27.5%
Informativeness	45.3%	29.2%	25.5%
Human-likeness	46.2%	29.3%	24.5%

Table 2: Human evaluation for DialoFlow and DialoGPT on the DailyDialog test Dataset.

such as "I don't know", which is a more suitable metric than BLEU when dealing with multi-reference test sets. We also use Entropy (Zhang et al., 2018) to evaluate the lexical diversity. We employ the evaluation scripts used by DialoGPT.

For interactive dialogue evaluation, we compute the Pearson and Spearman correlation between the automatic metrics and human ratings. We use the pre-trained DialoFlow-large to compute our proposed Flow score.

4 Results and Analysis

In this section, we show the performance of our pre-trained DialoFlow model on response generation as well as the performance of Flow score on interactive dialogue quality evaluation.

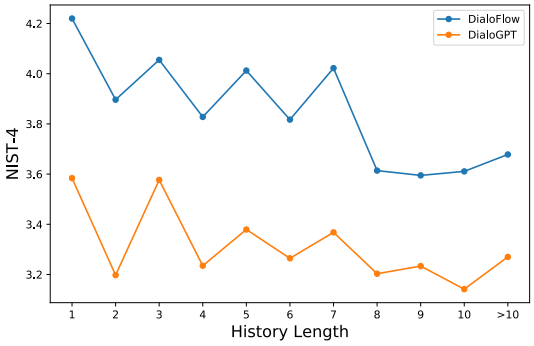

Figure 3: The performance of the large version of DialoFlow and DialoGPT on the samples of different history lengths on the DailyDialog dataset.

4.1 Response Generation

Table 1 lists the comparison of our pre-trained DialoFlow with the pre-trained DialoGPT on the Reddit multi-reference dataset. Generally, DialoFlow-large achieves the highest score on the NIST and METEOR, while DialoGPT-medium performs better on the BLEU. The performance of our DialoFlow increases with the model size, while the DialoGPT gets the best performance with the medium size rather than the large size. As NIST can effec-

Methods	B1	B2	B3	B4	B5	B6	B7	B8	B9	B10	B11	Human
Human ↑	4.142	4.140	4.075	4.035	3.933	3.864	3.849	3.848	3.828	3.692	3.605	*5.000*
FED ↑	4.988	4.818	4.621	4.670	4.555	4.739	4.438	4.355	4.651	4.799	3.608	3.468
Perplexity ↓	600.0	521.2	441.2	561.6	367.7	1731	1879	13347	662.2	618.4	50.29	51.39
Flow ↓	1.396	1.410	1.402	1.406	1.407	1.422	1.425	1.417	1.425	1.461	1.466	1.333

Table 3: The human ratings and automatic metrics for different chatbots. B1∼B11 denotes the 11 different chatbots in the DSTC9 Interactive Dialogue Evaluation Track. Human denotes the performance on the human-human conversations from the BST dataset. We assume the human rating for the human-human conversations is 5.000.

Method	Pearson	Spearman
FED	0.67 (p <0.1)	0.56 (p <0.1)
Perplexity	0.12 (p≈0.72)	0.20 (p≈0.55)
Flow	**0.91** (p <0.001)	**0.90** (p <0.001)

Table 4: Chatbot-level correlations on the DSTC9 Interactive Conversation dataset.

tively penalize common n-grams such as "I don't know", the results reveal that DialoGPT tends to generate general responses while our DialoFlow model can create more informative responses. The results also reflect that modeling the dynamic flow is helpful to boost the conversion quality and avoid converging to the general responses. For the lexical diversity, DialoFlow performs similarly with the DialoGPT on Entropy.

The average history length of the multi-reference Reddit dataset is only 1.45, which is a bit short. Thus, we conduct extensive experiments on the DailyDialog dataset (average history length = 4.66) to verify the performance gain on the long dialogue history. As shown in Table 1, DialoFlow shows significant improvements on all model sizes and on all metrics compared to the DialoGPT. The improvements on the DailyDialog dataset demonstrate that our DialoFlow model shows a great capacity to capture the dynamic information flow with a long history. Note that the performance improvement of the DailyDialog dataset is more remarkable than Reddit. In our opinion, conversations in Reddit are mainly the comments in forums, while in DailyDialog the dialogues are derived from daily life. Thus, in the DailyDialog dataset, the context flows are in the more similar schema, and the semantic influences are more predictable compared to the Reddit dataset.

Human Evaluation. We conduct human evaluation on 200 randomly sampled cases from the DailyDialog test dataset using crowd-sourcing. We compare DialoFlow and DialoGPT on the medium version. Each response pair is randomly presented to 3 judges, who rank them for relevance, informativeness, and human-likeness. The overall judge preferences are presented as a percentage of the total, as shown in Table 2. There is a strong preference for the responses generated by DialoFlow. The human evaluation demonstrates that modeling the dynamic information flow is effective for improving the quality of dialogue generation.

Analysis of dialogue history length. Figure 3 shows the performance of our DialoFlow and the DialoGPT on different history lengths. Overall, our DialoFlow achieves better performance on all history lengths. In particular, when history length equals 1, that is, the response is generated based on one history utterance, our DialoFlow also gains a prominent boosting. We attribute it to the guidance of predicted semantic inference.

Ablation Study. To explore the effect of the proposed training objectives, we conduct ablation studies on the medium version of DialoFlow, as shown in Table 1. With all three training objectives, DialoFlow model achives the best performance on NIST and METEOR. When we drop the Semantic Influence Modeling task, the performance slightly decreases. When we further drop the Context Flow Modeling task, which means the end-to-end training, the performance decreases again. The results reveal that the Context Influence Modeling task is effective for dialogue modeling and the Semantic Influence Modeling task can prompt the CIM task.

4.2 Dialogue Evaluation

Results. Table 4 shows the chatbot-level correlations of different automatic metrics with human ratings on the DSTC9 Interactive Conversation dataset. Our proposed Flow score achieves strong Spearman correlation of 0.90 (p<0.001) and strong Pearson correlation of 0.91 (p<0.001). FED only shows moderate correlations with a chatbot-level Spearman correlation of 0.56 (p<0.1). Perplexity score shows a very weak correlation. On the one hand, the results reveal that our proposed Flow score can effectively estimate the overall chatbot

Figure 4: The semantic context 2-D T-SNE visualization of a human-bot conversation. Our DialoFlow model captures the context flow, especially the topic changes. "0" is the start point. Better view in color.

quality. On the other hand, high correlation also demonstrates that the DialoFlow model captures the general dynamic information flow in the natural human-human conversation.

Results Analysis. Table 3 shows the detailed human ratings, FED scores, perplexity, and our proposed Flow score for the 11 chatbots in the DSTC9 Interactive Dialogue Evaluation Track and the sampled human-human conversations. Good automatic metrics should perform well not only on human-bot conversations but also human-human conversations because the ultimate goal of the chatbot is to generate human-like responses. FED performs poorly on the human-human conversations compared to its performance on the other 11 chatbots. Our proposed Flow score takes the human-human conversations as the best one, and the Flow score gap between human-human conversations and the best chatbot is similar to the human rating gap.

Analysis about Flow score. The Flow score can be regarded as the perplexity on the utterance level. There are many different expressions for a specific semantic in natural conversations. Traditional word-level perplexity can estimate the coherence and fluency of the utterance but always performs unstably on variable expressions. The Flow score directly measures the semantic similarity and alleviates the problem with the traditional perplexity.

4.3 Case Study

Figure 4 shows the 2-D T-SNE visualization of the semantic context of a human-bot conversation encoded by our pre-trained DialoFlow model. The conversation can be split into four topics: greetings (1∼4), talking about why bad day (5∼13), explaining the terrible experience seeing the doctor (14∼18), and discuss swimming (19∼26). Correspondingly, in the visualization, the semantic context flow visualization changes a lot when the topic switches, revealing that DialoFlow can capture the dynamic information flow in the dialogue and effectively measure the semantic influence brought about by each utterance. Besides, it seems like that different speakers keep their own context flows.

5 Related Works

Multi-turn dialogue modeling. The modeling of multi-turn dialogue history mainly falls into two categories: 1) Flat concatenation. These works directly concatenate the dialogue history as the input sequence (Zhang et al., 2020), which can not capture the information dynamics. 2) Hierarchical architectures. The hierarchical architecture is commonly used in the dialogue history understanding. Serban et al. (2016a) propose the hierarchical LSTM to generate responses. Li et al. (2019) introduce an incremental transformer to capture multi-turn dependencies. Shan et al. (2020); Gu et al. (2020) employ pre-trained BERT to encode individual utterances and design the utterance-

level encoder to capture the turn-level structure. These methods suffer from the lack of context word-level information when encoding utterances. Different from these methods, our DialoFlow takes full advantage of both word-level information and utterance-level dynamic information. Besides, the proposed DialoFlow is pre-trained on the large-scale open-domain dialogue dataset.

Pre-trained models for dialogue generation. Recent advances in pre-trained language models have great success in dialogue response generation. DialoGPT(Zhang et al., 2020), Plato-2 (Bao et al., 2020), Meena(Adiwardana et al., 2020), and Blender(Smith et al., 2020) achieve strong generation performances by training transformer-based language models on open-domain conversation corpus. In contrast, our proposed DialoFlow focuses on modeling the dynamic information flow in the pre-training process, and we design three training objectives to optimize the model.

Interactive Dialogue Evaluation. Evaluating the quality of interactive dialogue automatically is a challenging problem, as there is no gold reference for the utterances. Mehri and Eskénazi (2020) propose the FED score, an automatic dialogue evaluation metric using pre-trained DialoGPT-large, which works with pre-set common human comments, like *"It is interesting to talk with you."*, revealing the dialogue quality. However, the FED score has limited performance on those dialogues without apparent comments. Our Flow score entirely depends on the pre-trained DialoFlow model with no need for human integration.

6 Conclusion and Future work

In this work, we proposed the DialoFlow to model the dynamic information flow across dialogue utterances by addressing the semantic influence brought about by each utterance. Specifically, we employed a uni-directional Flow module to model the context flow and designed three training objectives to optimize the DialoFlow model. Besides, upon the DialoFlow, we proposed the Flow score, an automatic reference-free evaluation metric for interactive dialogue evaluation, with the pre-trained DialoFlow. Experiments on response generation and dialogue evaluation all demonstrate that our method could effectively capture the dynamic information flow across utterances. For future work, we would like to apply the DialoFlow to the task-oriented dialogue and explore the application on the long text generation, such as the story generation.

Acknowledgement

We sincerely thank the anonymous reviewers for their thorough reviewing and valuable suggestions. This work is supported by National Key R&D Program of China (NO. 2018AAA0102502).

References

Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a human-like open-domain chatbot. *CoRR*, abs/2001.09977.

Siqi Bao, Huang He, Fan Wang, Hua Wu, Haifeng Wang, Wenquan Wu, Zhen Guo, Zhibin Liu, and Xinchao Xu. 2020. PLATO-2: towards building an open-domain chatbot via curriculum learning. *CoRR*, abs/2006.16779.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020*, pages 830–839. AAAI Press.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Sarah Brown-Schmidt and Agnieszka E Konopka. 2015. Processes of incremental message planning during conversation. *Psychonomic bulletin & review*, 22(3):833–843.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2020. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. *CoRR*, abs/2012.01775.

Chulaka Gunasekara, Seokhwan Kim, Luis Fernando D'Haro, Abhinav Rastogi, Yun-Nung Chen, Mihail Eric, Behnam Hedayatnia, Karthik Gopalakrishnan, Yang Liu, Chao-Wei Huang, et al. 2021. Overview of the ninth dialog system technology challenge: Dstc9. *Proceedings of the 9th Dialog System Technology Challenge Workshop in AAAI2021*.

Prakhar Gupta, Shikib Mehri, Tiancheng Zhao, Amy Pavel, Maxine Eskénazi, and Jeffrey P. Bigham. 2019. Investigating evaluation of open-domain dialogue systems with human generated multiple references. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019*, pages 379–391. Association for Computational Linguistics.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation, WMT@ACL 2007, Prague, Czech Republic, June 23, 2007*, pages 228–231. Association for Computational Linguistics.

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, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995. Asian Federation of Natural Language Processing.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 12–21. Association for Computational Linguistics.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Shikib Mehri and Maxine Eskénazi. 2020. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 225–235. 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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 32–37. Association for Computational Linguistics.

Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016a. Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 588–598, Berlin, Germany. Association for Computational Linguistics.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016b. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784. AAAI Press.

Yong Shan, Zekang Li, Jinchao Zhang, Fandong Meng, Yang Feng, Cheng Niu, and Jie Zhou. 2020. A contextual hierarchical attention network with adaptive objective for dialogue state tracking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6322–6333. Association for Computational Linguistics.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

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, ACL 2020, Online, July 5-10, 2020*, pages 2021–2030. Association for Computational Linguistics.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 1815–1825.

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, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics.