Topic and Speaker-aware Hierarchical Encoder-Decoder Model for Dialogue Generation

Qitao Hu, Xiao Wei*
School of Computer Engineering and Science, *Shanghai University*, China
xwei@shu.edu.cn

***Abstract*—As one of the most common social behavior in human society, communication in multi-turn conversation or dialogue system has always been a research focuses of natural language processing (NLP). The quality of downstream tasks in multi-turn dialogue is often determined by the result of dialogue context modeling. For dialogue generation, the context information will determine the consistency and diversity of the generated responses. However, the current research on dialogue generation increasingly relies on external information rather than mining from the dialogue content itself. In this paper, we propose a topic and speaker-aware hierarchical encoder-decoder (TSHED) model to capture the topic and speaker information flow in the context for response generation with the hierarchical transformer-based framework. Specifically, we obtain semantic information of each utterance at word-level and then apply topic and speaker-aware attention to model context at utterance-level. Experimental results on two open-domain datasets show that TSHED significantly improves the quality of responses and outperforms strong baselines.**

***Index Terms*—open-domain dialogue system, dialogue generation, encoder-decoder model, transformer**

I. Introduction

Daily conversation is one of the crucial components of human social activities. Most daily conversations among groups of people can be summarized as open-domain dialogue or chit-chat, which are multi-turn and informative. From the perspective of dialogue content, the context is approximately equal to historical utterances in previous turns. Therefore, some studies focus on multi-turn conversations with context modeling for the downstream task of the dialogue system, such as dialogue generation, dialog emotion detection [1], abstractive dialogue summarization, etc. For dialogue generation, it is vital to make full use of context to generate consistent and logical responses. The earlier deep learning-based methods focus on short-text conversation or local context [2], [3]. Nevertheless, context modeling is not only a process of word alignment or sentence alignment, but also a process of perceiving the information flow that drives the dialogue. Serban et al. [4] proposed a ground-breaking hierarchical encoder-decoder framework called HRED to model context. HRED consists of a word-level encoder that encodes utterances, an utterance-level encoder that maps each utterance representation into dialogue context and a decoder that generates tokens. Since then, the ideas about context modeling with hierarchical architecture have been widely used in multi-turn dialogue generation tasks [5]–[7].

*Corresponding author: Xiao Wei, xwei@shu.edu.cn
DOI reference number: 10.18293/SEKE2023-046

However, compared with documents, the topic flow in dialogue

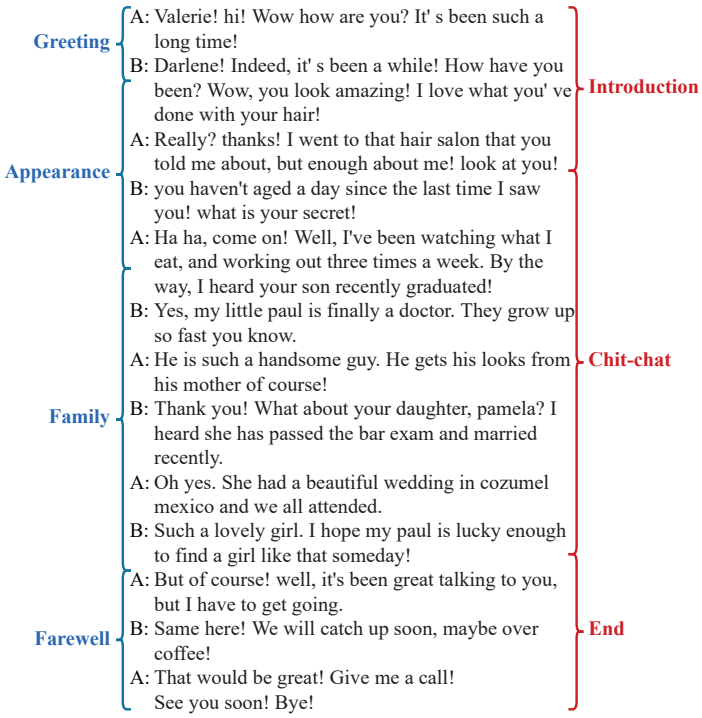

Fig. 1. A conversation from DailyDialog. The **blue** word represents dynamic topic word, the **red** word represents general topic word.

is more complex, which means the topic density of dialogue is higher and the topic shift in dialogue is more unpredictable. In addition, the speaker flow is unique to dialogue compared with other natural language scenes. The participants in the dialogue are at least more than one, which makes different interlocutors in the same scene with different response ideas. In our work, we introduce two granularity of topics, called dynamic topic which means the topic that arises as the conversation progresses and general topic which means the topic that reflects the phases of the dialogue process. As shown in Fig.1, this is the daily conversation between two middle-aged women who meet and chat. From the perspective of the topic, their topics after greeting are pretty random, such as 'appearance' and 'family', but their dialogue content follows the fixed pattern of greeting, chatting and saying goodbye. From the perspective of the speaker, 'B' shows a polite and warm attitude. The above instance shows that the topic and the

speaker are essential supplements to the context information in a multi-turn conversation scenario, affecting the quality of the generated response. In recent research on topic-aware response generation, the use of the topic includes extracting topic words or using external knowledge. Zhang et al. [8] took meaningful words that appeared in the dialogue history and response as topic words to generate topic-relevant responses. Feng et al. [9] used topic-aware attention and external knowledge sources to promote the understanding of dialogue history. In terms of the speaker, recent work uses external information about the speaker to model. Majumder et al. [10] used commonsense knowledge bases to expand and enrich persona descriptions.

However, there is still internal mining space in dialogue history rather than just external information to supplement the context in dialogue context modeling. In this paper, we propose a new model called TSHED to fully use the dialogue history with a multi-degree attention mechanism in context modeling for response generation. The core idea is applying several context-aware attention to model context at the utterance level. Specifically, we improve the traditional text segmentation algorithm TextTiling [11], complete the segmentation of two granularity of topics, and then use the transformer-based word-level encoder to get the representation of each utterance. After that, the utterance-level encoder obtains the representation of the whole context via topic-aware attention and speaker-aware attention. Finally, the GRU decoder is used to generate response token by token in combination with the attention weight. Experiments on two open-domain datasets demonstrate the effectiveness of TSHED, and further analyses reveal the advantage of TSHED in achieving state-of-the-art performances on the datasets.

II. Related Work

As one of the hot topics in natural language processing (NLP), dialogue systems which are highly coincident with human interaction have attracted much attention in recent years. In particular, dialogue generation in dialogue system is a challenging task, the research focus of which has gradually changed from single-turn to multi-turn. One reason is that multi-turn conversations are more common in daily life. Longer dialogue turns mean more complex context composed of speaker messages which determines conversation topics, speaker goals and style [5]. How to model context to generate better responses is a challenge in this area.

Earlier Studies often used representations of words in dialogue history accumulated to contextualize or focus on the recent context [2], [3]. Serban et al. [4] proposed Hierarchical Recurrent Encoder-Decoder (HRED), which built a ground-breaking hierarchical encoder-decoder framework to achieve context-awareness in dialogue systems. Since then, hierarchical-based models have been widely studied in the task of multi-turn dialogue generation. However, the information distribution of the context is uneven, and it is obviously inappropriate to treat all contexts equally. Therefore some researchers try to introduce the attention mechanism. Tian et al. [12] compared the non-hierarchical model with the hierarchical model. They proposed WSeq finding that neural networks can produce longer, more meaningful and diverse responses with more context information. Zhang et al. [13] introduced a self-attention mechanism to capture long-distance dependencies. With the extensive use of pre-trained language models in natural language generation tasks [14], it is possible to obtain richer semantic information via various pretraining objectives. Gu et al. [15] employed a hierarchical Transformer architecture with two training objectives to capture hierarchical coherences on dialogue generation tasks. However, the role of topic in dialogue context modeling had been almost ignored in the above research. In addition to implicit topic modeling with latent variables [5], [7], Yoshikoshi et al. [16] used keywords or topic entities based on dialogue history to explicit topic modeling. Nevertheless, the topic dominated by utterances may not be represented by variables or a word. Inspired by the above research work, we transfer dialogue history into information blocks from the perspective of topic and speaker, and then combine topic-aware and speaker-aware attention mechanisms into hierarchical encoder-decoder framework to model meaningful and informative context for the task of response generation.

III. Methodology

Our model consists of two parts: (1)Multi-granularity topic segmentation, which is used to segment the multi-turn dialogue history with topic boundaries in an unsupervised way. (2)TSHED Network, which uses topic and speaker-aware attention with hierarchical architecture to achieve context encoding, and finally decodes to generate responses.

A. Topic Segmentation

Compared with documents [17], dialogues are less structured, more organized and promoted by topics. We call topic flow hidden under the dialogue. In our paper, topic flow is not unique because topics are multi-granularity. For example, in a daily conversation among colleagues, the common topic flow of "greeting$\longrightarrow$work$\longrightarrow$psychology$\longrightarrow$vacation" can be abstracted from it. The trend of topic flow after the topic "vacation" can lead to any others, so we call this kind of topic flow in a dialogue as dynamic topic flow. Besides, some topics that represent the rhythm of conversation are common to most dialogues, such as "introduction" or "end" which means the beginning or end of a dialogue ; we call this kind of topic flow in a dialogue as general topic flow.

Such topic flow can be helpful in understanding the process of dialogue and the consistency of response generation. Given a continuous multi-turn dialogue $D = \{u_1, u_2, \ldots, u_n\}$, where n is the number of utterances, we combine classic unsupervised topic segment algorithm, TextTiling [11] that focus on the change of topic similarity in the text, with sentence represented by pretrained Sentence-BERT [18], to divide multi-granularity topic boundaries. The differences from traditional TextTiling are shown in the Fig. 2(b), first, we keep the integrity of the sentence instead of dividing the text unit according to the number of words with fixed length and

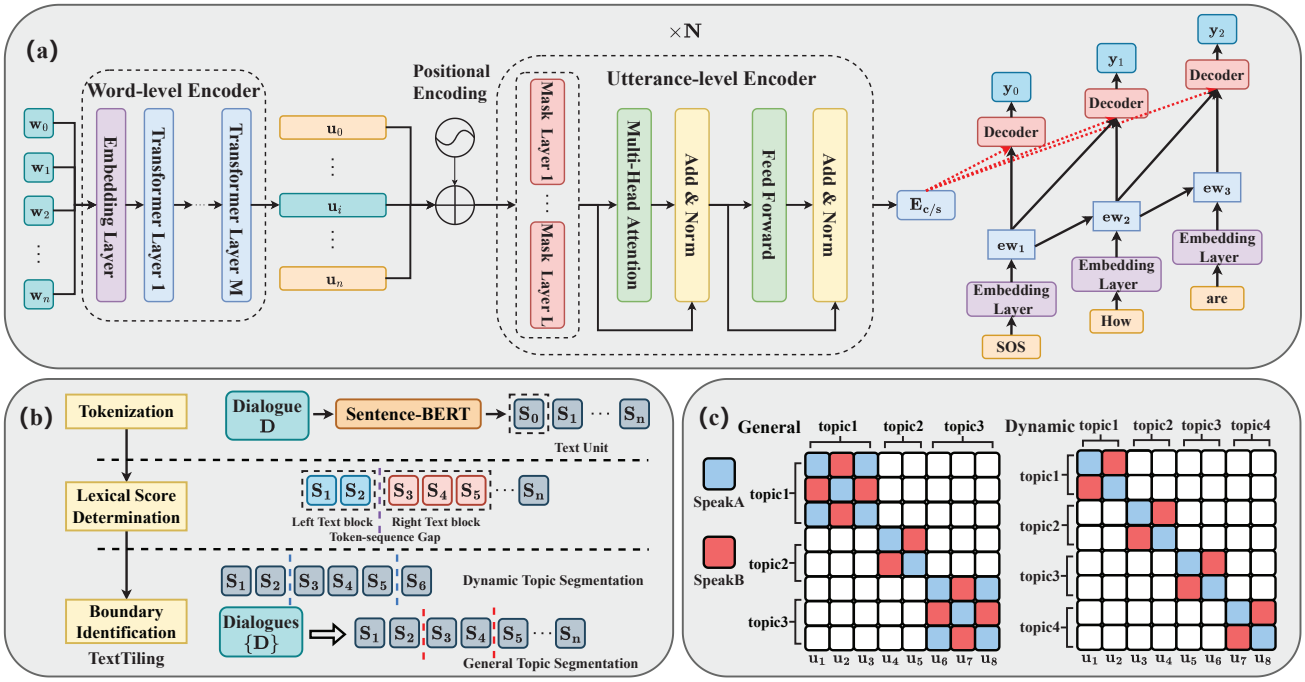

Fig. 2. Overview of our model. (a) The structure of TSHED, which is a hierarchical Transformer-based model with topic and speaker-aware attention. (b) The improvements on the three stages(tokenization, lexical score determination and boundary identification) of TextTiling. (c) The different masks are prepared for contextual information fusion. Due to layout size limitations, the representation of the speaker mask is mixed with that of the topic mask.

take the utterance as the text unit; second, we dynamically maintain the unbalanced length of text blocks(formed by several text units) on both sides of token-sequence gap to obtain a more realistic token-sequence gap; third, we segment multi-granularity topics with the perspective of dynamic and general topics.

To be specific, each utterance u_i is encoded as a sentence embedding se_i by Sentence-BERT, then the similarity sequence is obtained by calculating the similarity combination of blocks with different lengths, which can be used to convert into two sequences; one is used to calculate the depth score sequence for dynamic topic sequence denoted as $Topic_d = [t_1^a, t_2^a, t_3^b, t_4^b, t_5^b, \ldots, t_n^x]$, the other is used to calculate the compression similarity sequence for local general topic sequence denoted as $Topic_g^l = [t_1^1, t_2^1, t_3^2, t_4^2, t_5^2, \ldots, t_n^y]$, where t_i^s is the topic s of u_i and we call the gap between t_i and t_j as topic boundary, $y-1$ is the max number of general topic boundaries. Then global general topic sequence denoted as $Topic_g^g$ can be obtained by calculating the index average of each topic boundary in $Topic_g^l$, which is as general topic sequence $Topic_g$ finally.

B. TSHED Network

Compared with the document, there is another prominent feature in dialogue: more than one speaker participates in dialogue. Different speakers' rhythms or narrative characteristics make the text features of conversation diverse. Thus, we design a **T**opic and **S**peaker-aware **H**ierarchical **E**ncoder-**D**ecoder model **TSHED**, which encodes utterances from word-level to utterance-level by a hierarchical architecture as shown in Fig. 2(a). During utterance-level encoding, TSHED considers the segmentation of different speakers into context while referring to the multi-granularity topic mentioned above. For any multi-turn dialogue, we denote the dialogue as $D = \{u_1, u_2, \ldots, u_n\}$, the dynamic topic segmentation of D as $Topic_d$, the general topic segmentation of D as $S_{gt} = Topic_g$, the speaker segmentation of D as $S_{sp} = [s_1^1, s_2^2, s_3^1, \ldots, s_n^2]$.

1) Word-level Encoder: The word-level encoder we used in TSHED is transformer-based, designed to extract the semantic information of each utterance in the dialogue history D. For each word, we use the fixed dimension of the trainable embedding matrix E_w to represent and the embedding ew_{ij} corresponding to the i-th word in the j-th utterance. Then for each word embedding, we use the standard transformer encoder to encode, by taking the contextual embedding of the last token of each utterance (In our model, the special token is $\langle eou \rangle$) to represent every utterance individually. We denote the output of the word-level Encoder:

$$\begin{aligned} E_u &= \{eu_1, eu_2, \ldots, eu_m\} \\ &= \text{WEncoder}(u_1, u_2, \ldots, u_m) \end{aligned} \tag{1}$$

where $u_j = \{w_{1j}, \ldots, w_{nj}\}$ is the j-th utterance.

2) Utterance-level Encoder: In a multi-turn dialogue, topics of different granularity are usually threaded through to drive the conversation. In addition, different speakers with different dialogue rhythms both play a dominant or passive role in the process of dialogue. Therefore, the topic and speaker serve as crucial contextual information. The utterance-level encoder in our model inspired by Transformer [19] focus

on context-aware modeling, processing all word-level encoded utterance in the dialogue history from the perspective of the topic and speaker to obtain the semantic information among utterances. To be specific, S_{dt}, S_{gt}, S_{sp} as fixed-length vectors with the same length as the dialogue history, are fed into utterance-level encoder with sentence embedding. Then, in addition to the two granularity of topics, in order to further capture the relationship between the topic and the speaker, we introduce the heterogeneous multi-head self-attention with three kinds of masks, which can be formulated as:

$$\text{Attention}(Q, K, V, M) = \underset{\text{seq}}{\text{softmax}}\left(\frac{QK^\top}{\sqrt{d_k}} + M\right)V \quad (2)$$

where M is the mask matrix, determines whether the utterance can be attended to. Specifically, according to S_{dt}, S_{gt}, S_{sp}, we introduce several segmentation masks M^{type} as shown in Fig. 2(c), which can be defined as:

$$\begin{aligned} M_{ij}^{dt} &= \begin{cases} 0, & \text{same dynamic topic} \\ -\infty, & \text{otherwise} \end{cases} \\ M_{ij}^{gt} &= \begin{cases} 0, & \text{same general topic} \\ -\infty, & \text{otherwise} \end{cases} \\ M_{ij}^{sp} &= \begin{cases} 0, & \text{same speaker} \\ -\infty, & \text{otherwise} \end{cases} \end{aligned}$$

where i, j is the position of the utterance in the dialogue. Each head in multi-head self-attention mechanism is defined as:

$$\text{head}_i = \text{Attention}\left(E_u W_i^Q, E_u W_i^K, E_u W_i^V, M\right) \quad (3)$$

where $W_i^Q \in \mathbb{R}^{dmodel \times d_q}$, $W_i^K \in \mathbb{R}^{dmodel \times d_k}$, $W_i^V \in \mathbb{R}^{dmodel \times d_v}$ are learnable parameter matrices, d_q, d_k, d_v denote the dimension of Query vectors, Key vectors and Value vectors. Each head focuses on different contextual information and the multiple contextual information is fused through concatenation, which can be formulated as:

$$\begin{aligned} E^c &= \text{UEncoder}(eu_1, eu_2, \ldots, eu_m) \\ &= \text{Concat}\left(\text{Head}^{base}, \text{Head}^{segt}\right) W^O \\ \text{Head}^{base} &= \text{Concat}\left(\text{head}_1, \ldots, \text{head}_x\right) \\ \text{Head}^{segt} &= \text{Concat}\left(\text{head}_1, \ldots, \text{head}_y\right) \end{aligned} \quad (4)$$

where $W^O \in \mathbb{R}^{hd_v \times d_{model}}$, h denotes the total number of heads and $x + y = h$. Head^{base} denotes the set of heads without segmentation masks and Head^{segt} denotes the set of heads with three kinds of segmentation masks. E^c denotes the context representation.

3) Decoder: The response is generated by a recurrent decoder token by token with attention.

$$\begin{aligned} \mathbf{h}_t^{\text{dec}} &= \text{BiGRU}\left(\mathbf{h}_{t-1}^{\text{dec}}, [\mathbf{e}_{t-1}^w; \mathbf{C}_t; \mathbf{S}_t]\right) \\ \mathbf{C}_t &= \text{attention}(\mathbf{h}_{t-1}^{\text{dec}}, E^c) \\ \mathbf{S}_t &= \text{attention}(\mathbf{h}_{t-1}^{\text{dec}}, E^S) \end{aligned} \quad (5)$$

where $\mathbf{h}_t^{\text{dec}}$ denotes the hidden state at the t step, $\mathbf{e}_{t-1}^w$ denotes the word embedding of a generated word at the $t-1$ step, $\mathbf{C_t}$ and $\mathbf{S_t}$ denotes the representation of context and the same speaker as the respondent with additive attention [20] weight.

$$y_{t+1} = \text{Softmax}(\mathbf{W}_{\text{dec}}\mathbf{h}_t^{\text{dec}}) \quad (6)$$

where y_{t+1} denotes the generated word, $\mathbf{W}_{\text{dec}}$ is a matrix that aligns $\mathbf{h}_t^{\text{dec}}$ dimension with the dimension of the target vocabulary.

IV. EXPERIMENTS

A. Datasets

We evaluate our model on the following open-domain datasets:

DailyDialog [21] is a high-quality, multi-turn dialog dataset, the manually labeled dialogs of which are human-written to reflect our daily conversations.

Cornell Movie Dialog Corpus [22] is a dialog dataset of fictional conversations extracted from raw movie scripts.

For each dataset, we split the corpus into the training set, validation set and test set at the ratio of 8:1:1, and reserve the dialogue containing more than 3 utterances. More details of the two datasets are shown in TABLE I.

TABLE I
STATISTICS OF THE DATASETS.

Dataset	DailyDialog	Cornell Movie Dialog Corpus
dialogs number	64190	93513
vocabulary	18091	42315
turn length	6.52	6.21
utterance length	14.54	11.62

B. Implementation Details

Our model is implemented using Pytorch. For topic segmentation, we load the pre-trained "all-mpnet-base-v2" for Sentence-BERT to get representations of each utterance. At the stage of topic segmentation, for obtaining dynamic topic boundary, we set the max length of the text block 7; for obtaining general topic boundary, we set the max number of the text block 3. At the stage of response generation, the numbers of hidden nodes are all set to 300, and the encoder and decoder layers are set to 10, 8. We set the number of heads in the multi-head mechanism of the utterance-level encoder as 12 and the ratio of the $head_{base}$, $head_{dtopic}$, $head_{gtopic}$ and $head_{speaker}$ is $4:2:1:1$. The decoding strategy we use is top-k sampling and nucleus sampling. During training, we set the batch sizes to 32 and 16 for DailyDialog and Cornell Movie Dialog Corpus datasets, respectively. The initial learning rate we set to 0.0001. Adam is used as our optimizer. We train our models at least 100 epochs on RTX 2080Ti GPUS.

C. Baselines

We compare our model with the following baselines with end-to-end framework on response generation task:

Seq2Seq [23] is a model with attention mechanism uses the sequence-to-sequence framework.

TABLE II
EXPERIMENTAL RESULTS ON DAILYDIALOG AND CORNELL MOVIE DIALOG CORPUS WITH AUTOMATIC EVALUATIONS.

Model	Average	Extrema	Greedy	Distinct-1	Distinct-2	BERTScore
Dataset: DailyDialog						
Seq2Seq	0.596	0.759	0.488	0.014	0.069	0.212
VHRED	0.611	0.773	0.497	0.028	0.129	0.214
HRAN	0.634	0.783	**0.527**	0.027	0.158	0.238
ReCoSa	0.615	0.770	0.497	**0.031**	0.157	0.216
TSHED	**0.641**	**0.798**	0.523	0.029	**0.164**	**0.339**
Dataset: Cornell Movie Dialog Corpus						
Seq2Seq	0.529	0.625	0.439	0.006	0.030	0.121
VHRED	0.531	0.632	0.442	**0.008**	0.039	0.126
HRAN	0.522	0.623	0.438	0.007	0.049	0.127
ReCoSa	0.521	0.620	0.436	0.007	0.032	0.121
TSHED	**0.548**	**0.633**	**0.455**	0.006	**0.052**	**0.129**

VHRED [5] is a HRED-based model combined a latent variable into generation process.

HRAN [6] is a hierarchical attention framework model focusing on the important information in context.

ReCoSa [13] is a dialogue generative model which makes full use of self-attention mechanism to find the relevant contexts.

D. Metrics

We adopt the following automatic metrics to evaluate the generated response:

Average, Extrema, Greedy [5] are the embedding-based metrics to measure the semantic similarity to the ground-truth response. We use the pre-trained Word2Vec word embeddings on the Google News Corpus for evaluation.

Distinct-n [3] is the metrics for reporting the degree of diversity, which is defined as the ratio of unique uni/bigrams to the total number of uni/bigrams in generated responses.

BERTScore [24] is the diversity metric used in the task of text generation, which computes a similarity score between two sentences that uses pre-trained BERT feature extraction.

E. Results and Analysis

TABLE II shows the results on two datasets. Generally, the proposed TSHED outperforms other models in most metrics on both datasets. As a unique non-hierarchical architecture in the experiment, Seq2Seq performs the worst. As the attention-based models, HRAN, ReCoSa, TSHED outperform VHRED on DailyDialog dataset. But on Cornell Movie Dialog Corpus dataset, only TSHED outperform VHRED. One reason for the result is that the dialogue turn and utterance length of DailyDialog dataset are longer than that of Cornell Movie Dialog Corpus dataset. The carefully designed attention mechanism in models can be fully utilized. Cornell Movie Dialog Corpus dataset is small in size of a dialogue, but very diverse and complex in content and style [25]. In addition, Cornell Movie Dialog Corpus dataset contains conversations from movies, and the character feature is more distinct, so the speaker-aware TSHED outperforms.

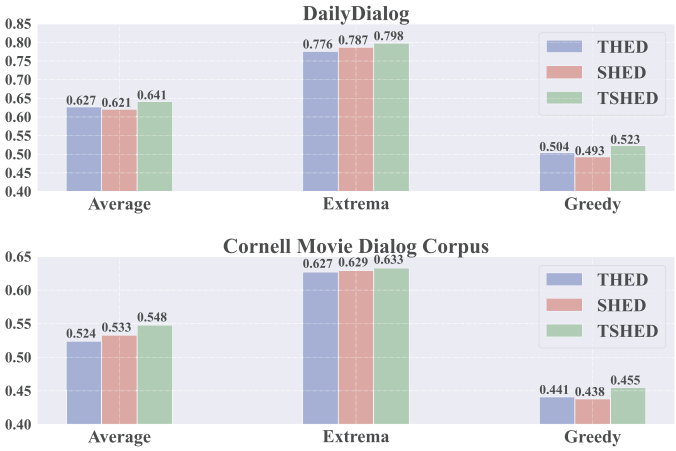

Fig. 3. The results of models with different blocks on the datasets.

F. Ablation Study

In order to better understand the impact of the different context-aware attention mechanisms in our model, we have conducted experiments with different attention-based utterance-level encoders on the datasets. We denote TSHED without topic-aware mask as SHED, TSHED without speaker-aware mask as THED. As shown in Fig.3, the performance of THED and SHED is different on different datasets, but TSHED achieves better performance than T/SHED, demonstrating that all of our designed masks which is related to information flow in a dialogue are critical to context modeling.

G. Case Study

To better illustrate the coherent response TSHED generates in multi-turn dialogue. we provide a long multi-turn(12 turns) dialogues in TABLE III from the DailyDialog dataset. Seq2Seq, VHRED, HRAN tend to generate general and emotional responses which are not in line with the background of business negotiation. Compared with the response of ReCoSa, that of TSHED is more coherent and human-like. Note that the longer a dialogue, the more information about topics, scenes and speakers. From the case above, we find that with the help of explicit prompt in context, responses generated in multi-turn dialogues are not only relevant but also meaningful.

TABLE III
CASE STUDY FOR MULTI-TURN DIALOGUE GENERATION.

History	... B: Mmmm, sounds interesting, gourmet chocolates, where are they produced? Belgium? A: Actually, the factory is located in scotland. B: Really? I didn't think they were known for their luxury chocolate production. A: That's what makes this such a fantastic opportunity! the government is one hundred percent supportive of creating new export markets and has guaranteed a low tariff for all wholesale orders of over one thousand units. They've also reduced the red tape involved at customs as well. Here, I brought these especially for you, try one! B: Oh, thanks. Mmm, hmm, creamy texture, very smooth ... A: Unique, aren't they? I bet you've never tasted anything like it! Quality is assured as I personally visit the factory to make sure no one' s cutting corners with the ingredients. Only the creme make it through inspection. B: Yes, very interesting flavors ... slightly spicy, very unique, that's for sure. Exactly what are the ingredients? A: I have it on highest authority that this traditional secret recipe has been handed down in the Granger family for generations. I'm sure you can keep a secret. Buttermilk, cacao beans, sugar and haggis.
GT	**Haggis? What's haggis?**
Seq2Seq	Oh, I see! That' s great! I' m so happy with it. I' m gonna have.
VHRED	Wow, that 's great! I' m glad. That' s gonna be great!
HRAN	That's a good idea! I' m gonna have a look at the same as a matter of fact!
ReCoSa	How do you know the secret?
TSHED	**Oh, wonderful! What' s the haggis?**

V. CONCLUSION AND FUTURE WORK

In this paper, we propose a topic and speaker-aware hierarchical encoder-decoder (TSHED) model to enrich informative context for response generation in multi-turn conversations. We apply a two-stage strategy to obtain topic boundaries and then adopt a hierarchical framework to achieve context-aware modeling. Experimental results show that our model performs well on DailyDialog and Cornell Movie Dialog Corpus datasets.

In future work, we plan to improve the effect of downstream tasks in multi-turn dialogue by combining the two stages of topic segmentation and dialogue context encoding or introducing emotion detection into word-level encoder.

REFERENCES

[1] X. Wang, L. Kou, V. Sugumaran, X. Luo, and H. Zhang, "Emotion correlation mining through deep learning models on natural language text," *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4400–4413, 2021.

[2] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," in *Proc. of ACL-IJCNLP*, 2015, pp. 1577–1586.

[3] J. Li, M. Galley, C. Brockett, J. Gao, and W. B. Dolan, "A diversity-promoting objective function for neural conversation models," in *Proc. of NAACL-HLT*, 2016, pp. 110–119.

[4] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proc. of AAAI*, vol. 30, no. 1, 2016.

[5] I. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proc. of AAAI*, vol. 31, no. 1, 2017.

[6] C. Xing, Y. Wu, W. Wu, Y. Huang, and M. Zhou, "Hierarchical recurrent attention network for response generation," in *Proc. of AAAI*, vol. 32, no. 1, 2018.

[7] L. Shen, Y. Feng, and H. Zhan, "Modeling semantic relationship in multi-turn conversations with hierarchical latent variables," in *Proc. of ACL*, 2019, pp. 5497–5502.

[8] S. Zhang, T. Zhao, and T. Kawahara, "Topic-relevant response generation using optimal transport for an open-domain dialog system," in *Proc. of COLING*, 2020, pp. 4067–4077.

[9] Y. Feng, G. Lampouras, and I. Iacobacci, "Topic-aware response generation in task-oriented dialogue with unstructured knowledge access," *arXiv preprint arXiv:2212.05373*, 2022.

[10] B. P. Majumder, H. Jhamtani, T. Berg-Kirkpatrick, and J. McAuley, "Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions," in *Proc. of EMNLP*, 2020, pp. 9194–9206.

[11] M. A. Hearst, "Texttiling: Segmenting text into multi-paragraph subtopic passages," *Computational linguistics*, vol. 23, no. 1, pp. 33–64, 1997.

[12] Z. Tian, R. Yan, L. Mou, Y. Song, Y. Feng, and D. Zhao, "How to make context more useful? an empirical study on context-aware neural conversational models." in *Proc. of ACL*, 2017, pp. 231–236.

[13] H. Zhang, Y. Lan, L. Pang, J. Guo, and X. Cheng, "Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation," in *Proc. of ACL*, 2019, pp. 3721–3730.

[14] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. of ACL*, 2020, pp. 7871–7880.

[15] X. Gu, K. M. Yoo, and J.-W. Ha, "Dialogbert: Discourse-aware response generation via learning to recover and rank utterances," in *Proc. of AAAI*, vol. 35, no. 14, 2021, pp. 12 911–12 919.

[16] T. Yoshikoshi, H. Atarashi, T. Kodama, and S. Kurohashi, "Explicit use of topicality in dialogue response generation," in *Proc. of NAACL-HLT*, 2022, pp. 222–228.

[17] X. Wang, H. Zhang, and Z. Xu, "Public sentiments analysis based on fuzzy logic for text," *International Journal of Software Engineering and Knowledge Engineering*, vol. 26, no. 9-10, pp. 1341–1360, 2016.

[18] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proc. of EMNLP-IJCNLP*, 2019, pp. 3982–3992.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. of NIPS*, 2017, pp. 5998–6008.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[21] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proc. of IJCNLP*, 2017, pp. 986–995.

[22] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," in *Proc. of ACL*, 2011, pp. 76–87.

[23] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[24] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," *arXiv preprint arXiv:1904.09675*, 2019.

[25] Y. Park, J. Cho, and G. Kim, "A hierarchical latent structure for variational conversation modeling," in *Proc. of NAACL-HLT*, 2018, pp. 1792–1801.