Learning to Improve Persona Consistency in Multi-party Dialogue Generation via Text Knowledge Enhancement

Dongshi Ju, Shi Feng*, Pengcheng Lv, Daling Wang , Yifei Zhang
Northeastern University, Shenyang, China
judongshi@163.com, fengshi@cse.neu.edu.cn
ricardolvpc@foxmail.com
{wangdaling, zhangyifei}@cse.neu.edu.cn

Abstract

In an open-domain dialogue system, the consistent persona is a key factor to generate real and coherent dialogues. Existing methods suffer from the incomprehensive persona tags that have unique and obscure meanings to describe human's personality. Besides, the addressee information, which is closely related to express personality in multi-party dialogues, has been neglected. In this paper, we construct a multi-party personalized dialogue dataset and propose a graph convolution network model (PersonaTKG) with addressee selecting mechanism that integrates personas, dialogue utterances, and external text knowledge in a unified graph. Extensive experiments have shown that PersonaTKG outperforms the baselines by large margins and effectively improves persona consistency in the generated responses.

1 Introduction

Endowing a dialogue agent with a consistent persona has attracted an increasing amount of research interests, as it helps to deliver a more coherent and engaging conversation for users. Existing studies explore character personality through key-value persona pairs (Qian et al., 2018) or short descriptive sentences (Zhang et al., 2018b). The key-value pairs define a few of persona categories, such as name, gender, and age, which have clear semantic meanings. The descriptive text is declarative sentences with relatively fixed patterns that introduce one's persona information such as occupation and hobbies in the first person. Promising performance have been achieved on these 'well-defined' persona datasets (Mohapatra et al., 2021; Gu et al., 2021a; Song et al., 2020b).

Recently, Li et al. (2020) proposed a new dialogue dataset HLA-Chat using tags as persona information. HLA-Chat was collected from scripts of hit TV dramas, and the persona tags of the characters were tropes that are determined by audiences' impressions on TVTropes website[1]. For example, the famous character *Sheldon Cooper* from *The Big Bang Theory* has the persona tags *Book Dumb*, *Omnidisciplinary Scientist*, *Green Eyed Epiphany*, *Neat Freak*, *Token Minority*, etc. We can observe that these tags (i.e. tropes generated by TV audiences) are usually very distinctive and rare words, as they represent the unique persona of the character. Different from the general and comprehensible persona definition in Qian et al. (2018); Zhang et al. (2018b), the tags in HLA-Chat contain rich persona information but are difficult to understand, which set obstacles for the model to generate persona consistent responses. We also argue that this incomprehensible persona challenge is different from generating the response from sparse persona data (Zheng et al., 2020), where there are only limited personalized sentences in the dialogue context.

A persona of Sheldon Cooper	
tag	Omnidisciplinary Scientist
sent	A scientist who knows everything about science.
doc	Related to the Nerd and the Mad Scientist, the Omnidisciplinary Scientist is a master of every branch of science, regardless of the branch in which they theoretically have a degree ...
Conversations	
Penny	I believe that when one door closes, another opens.
Sheldon Cooper	No,it doesn't. Not unless the two doors are conected by relays, or there are motion sensors involved.

Table 1: An example of persona tags with laconic and detailed interpretation and conversations involving the persona. The laconic interpretation (dubbed as sent) consists of one sentence, while the detailed interpretation is a long document that explicates the tag meaning.

Intuitively, the persona consistency in the generated responses of HLA-Chat can be further improved by incorporating external knowledge. However, most of these rare persona words could not be found in the knowledge base or commonsense base such as ConceptNet (Speer et al., 2017) and ATOMIC (Sap et al., 2019). Thanks to TVTRopes, the audiences can contribute to the laconic and de-

*Corresponding author

[1]https://tvtropes.org/

Proceedings of the 29th International Conference on Computational Linguistics, pages 298–309
October 12–17, 2022.

tailed interpretation of these tropes on this wiki website, as shown in Table 1. In this paper, we collect the user generated interpretation on TVTRopes as external knowledge, and introduce text-based knowledge in the persona consistent response generation model with unique and incomprehensive persona tags.

On the other hand, most of the existing personalized response generation studies focus on bilateral dialogue (Wu et al., 2021; Majumder et al., 2021). In effect, the conversations in the real world often occurs between multiple speakers, which is called multi-party conversation (MPC). MPC is generally composed of speakers, utterances, and addressees (i.e. the recipient corresponding to an utterance). The expression of persona in the dialogue is often closely related to the addressee. For example, the addressee can be a friend, a lover, a stranger, etc. For these different relationships, every speaker may have a different way of expressing their personas, so predicting the addressee can assist the dialogue system to promote the persona consistency in the responses. Thus, we incorporate the addressee selecting mechanism into the response generation model and leverage a posterior selection module to improve the addressee prediction.

The complex context structure of MPC urges researchers to constantly seek novel and effective context modeling methods, where graph convolution networks have already achieved promising results (Liang et al., 2021; Hu et al., 2019). However, none of the existing hierarchy (Meng et al., 2018), role sensitive (Liu et al., 2019), or GCN-based modeling methods consider the different personas of the speakers as well as the correlation between personas and utterances. In this paper, we first utilize the hierarchical recurrent encoder-decoder structure (Serban et al., 2016; Xing et al., 2018) and the bidirectional GRU (Penghua and Dingyi, 2019) to model the multi-turn dialogue context and the personas of all the speakers. Then we build a unified graph with utterances and personas as nodes, and employ GCN to aggregate dialogue context information.

Our contributions are summarized as follows:

(i) We construct a new personalized dialogue dataset HLA-Chat++[2], where each incomprehensive persona tag has laconic and detailed text-based interpretations.

(ii) We propose a **Persona**-consistent response generation model based on **T**ext **K**nowledge enhanced **G**CN (PersonaTKG) with addressee selecting mechanism that integrates personas, dialogue utterances, and external text knowledge in a unified graph. To the best of our knowledge, this is the first study that explores the addressee selection in personalized dialogue generation tasks.

(iii) We conduct extensive experiments on HLA-Chat++ dataset, and the results have validated the effectiveness of incorporating text knowledge and addressee selection in improving persona consistency of generated responses.

2 Related Work

2.1 Persona Consistent Dialogue Generation

Maintaining persona consistency is essential to delivering more realistic and coherent conversations. To incorporate persona information into the dialogue system, Li et al. (2016) first used persona embedding to project each speaker into a dense vector. Kottur et al. (2017) proposed a neural dialogue model that simultaneously considers the contextual history of the speakers. However, these two models rely heavily on data with persona annotation, which are expensive and sparse. Qian et al. (2018) defined multiple key-value pairs to represent the personas of the speakers, including information such as name, gender, age, and residence, and explicitly displayed these values in response. Zhang et al. (2018b) constructed PERSONA-CHAT dataset and proposed to model persona information using memory networks. On PERSONA-CHAT dataset, Yavuz et al. (2019); Song et al. (2019) explored the effectiveness of copy mechanism and conditional variational encoder. Further researches are conducted to promote the consistency of personas, such as a generation network based on personas to guide knowledge selection (Lian et al., 2019), a transmitter-receiver framework to explicitly model the understanding between speakers (Liu et al., 2020), and a multi-stage dialogue response generation framework to delete the words that may lead to inconsistency in the response, then rewrite it (Song et al., 2020a) on this basis. Majumder et al. (2020) adopted common sense databases and interpretation resources to expand persona information. Although these models have achieved promising results, they all have limitations when the persona information is incomprehensible, and incapable to learn persona

[2]https://github.com/NEU-DataMining/HLA-ChatPlusPlus

consistent expression effectively.

2.2 Multi-party Dialogue Generation

Existing methods of building dialogue systems can be generally categorized into studying two-party conversations and multi-party conversations. However, the task scenario of the multi-party dialogue system is closer to that in real life. In addition to predicting response, selecting the addressee of an utterance is also an important task for MPC. Ouchi and Tsuboi (2016) first proposed the task of addressee and response selection, and Zhang et al. (2018a) have validated the effectiveness of jointly modeling addressee and response selection. Le et al. (2019) proposed the who to whom (W2W) model to solve the problem of missing and completing addressee in a dialogue history. Tan et al. (2019) proposed the Context-Aware Thread Detection (CATD) to address the consistency of context and input messages. For response generation, (Hu et al., 2019) firstly tried to use a graph to model multi-party dialogue history, and effectively used the dialogue structure information. Liu et al. (2019) proposed interlocutor aware contexts into recurrent encoder-decoder (ICRED) frameworks model, which used three role GRUs to update the speaker vector, and then used the speaker vector and addressee information to generate a response. Wang et al. (2020) proposed to select responses accurately based on tracking dynamic topic. Gururangan et al. (2020); Gu et al. (2021b) adopted a multi-task learning method in MPC, which proves the effectiveness of incorporating domain knowledge. However, previous MPC researches have not studied persona information of the speakers.

3 Dataset Construction

Film and television drama scripts are a common dataset source for dialogue system research, where there are high-quality personas, distinctive characters, and many rounds of dialogues. We collects 30 English scripts from the website[3] to construct a multi-party dialogue dataset. Inspired by Li et al. (2020), we employ the tropes on TVTRopes as the personas of the characters, which are annotated by the audiences with more representative and distinctive meanings.

For the crawled script web pages, the main preprocessing steps are as follows: (i) Use regular expressions to filter out HTML escape characters and non-dialogue contents such as scenes, narration, and background of the script, and then generate the original dialogue dataset according to each scene of the script; (ii) Split and screen out the speakers in the script to build a multi-party dialogue dataset; (iii) Associate the characters in the dataset with the tropes in TVTRopes as personas. The supporting characters with no persona information are deleted. We collect the *Laconic* and *Main* user generated interpretation of tropes on TVTRopes Namespce[4] as sentence-level and document-level text knowledge. Sentence-level (*Laconic*) knowledge briefly explains the persona tags in one sentence, and document-level (*Main*) knowledge further explains the persona tags in detail through examples, as shown in Appendix A.

Finally, we construct a new multi-party dialogue dataset, HLA-Chat++, which has 823,204 conversations with character persona annotations. According to the statistics, HLA-Chat++ has 239 characters in the dataset, with an average of 3,444 dialogues per character, and an average of 27,440 dialogues per TV dramas. Table 2 shows the example TV dramas with data size and the number of main characters. Only 8 TV dramas are selected due to the space limitation.

TV dramas	size	characters
Alias	19,312	9
Bones	42,952	8
Charmed	25,572	8
Friends	23,520	7
GilmoreGirls	105,303	19
Merlin	13,822	6
NCIS	33,900	9
QueerAsFolk	14,123	7

Table 2: The multi-party personalized dialogue dataset

	HLA-Chat++	HLA-Chat
number of dramas	30	38
data size	823,204	1,042,647
persona source	TVTropes	TVTropes
persona representation	persona tags	persona sentences
number of sentence-level knowledge	4,778	-
number of document-level knowledge	4,778	-
average length of sentence-level knowledge	7.5	-
average length of document-level knowledge	487.7	-

Table 3: The multi-party personalized dialogue dataset

[3] http://transcripts.foreverdreaming.org/

[4] https://tvtropes.org/pmwiki/index_report.php

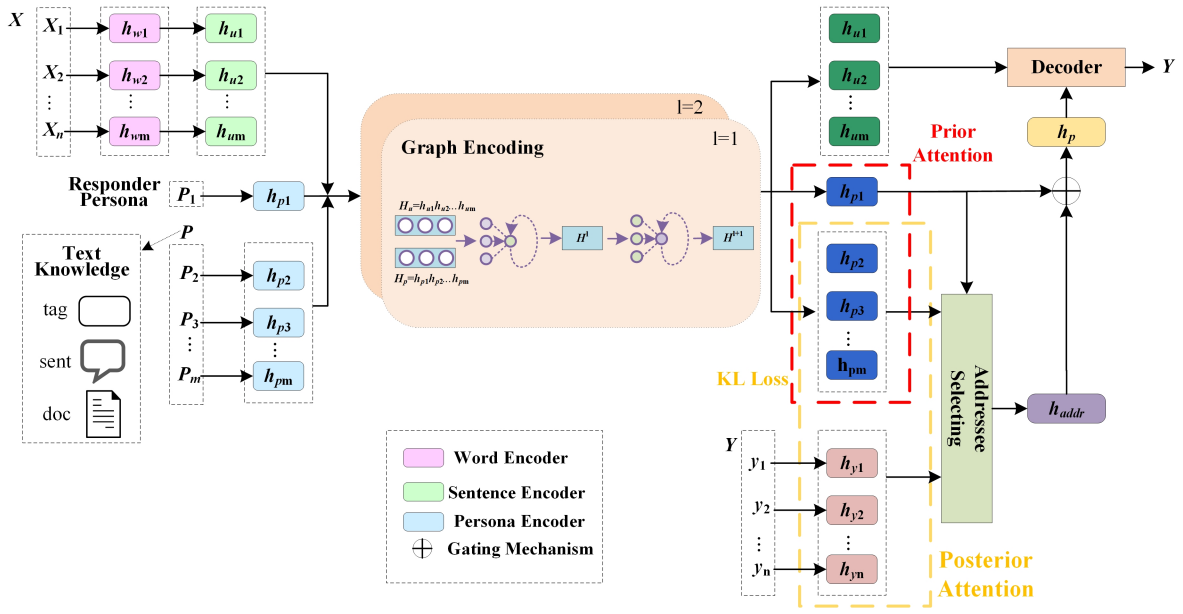

Figure 1: Framework of PersonaTKG

Compared with the original HLA-Chat, HLA-Chat++ has a smaller number of dialogues, as shown in Table 3. Li et al. (2020) processed the persona tags in HLA-Chat into descriptive sentences in the first person, just following the format in Zhang et al. (2018b). On the contrary, we retrain the persona tags, and collect the corresponding sentence-level and document-level interpretation of the tags, which can enrich the semantic meanings of these incomprehensible personas. Besides, HLA-Chat++ pays more emphasis on the multi-party dialogue structure, and preserves the relevant information in the dataset.

4 Model

PersonaTKG is shown in Figure 1, which is based on a Seq2Seq structure. The task can be formally defined as: given context $X = \{X_1, X_2, \ldots, X_m\}$, where X_i denotes the utterance of the speaker, and persona set $P = \{P_1, P_2, \ldots, P_k\}$, where k denotes k speakers, P_1 is the personas of the responder and $P_i (i > 1)$ is the personas of other speakers. The goal is to generate response $Y = y_1 y_2 \ldots y_n$ based on context and persona set, where y_i denotes the word generated in each step.

4.1 Context and Persona Encoder

In order to fully capture the information in multi-turn context, we adopt a hierarchical encoding strategy. The utterance encoder consists of word-level and sentence-level encoders, both of which are single-layer bidirectional GRU. The original data of the utterance encoder is X_i, the word-level encoder is responsible for encoding the utterances into vectorized representation $H_w = \{h_{w1}, h_{w2}, \ldots, h_{wm}\}$, which is calculated again by the sentence-level encoder, then the output of forward GRU and backward GRU are concatenated as the representation of X_i, namely $H_u = \{h_{u1}, h_{u2}, \ldots, h_{um}\}$, each h_{ui} contains the text information of X_i and the information immediately before and after X_i.

The original data of the persona encoder is the persona P_i of the speaker, which is encoded by single-layer bidirectional GRU to obtain the representation $H_p = \{h_{p1}, h_{p2}, \ldots, h_{pk}\}$. The representation of personas of the speakers is mainly used for subsequent addressee selection, considering that one can not accurately judge the addressee if it only contains persona information. If the personas of the speakers embed contextual information, it may be helpful for addressee prediction. Therefore, a graph convolution network is needed to further encode the representation of utterances and personas to aggregate information.

4.2 Graph Construction

The graph is represented as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which is contracted from the utterances and persons in the following way.

Vertexes: including utterance nodes and persona nodes, and then utterance and persona in the dialogue are concatenated to represent a vertex $v_i \in \mathcal{V}$

in $\mathcal{G}$, and each vertex v_i is initialized with the cooresponding sequentially encoded feature vector $H = \{h_{u1}, h_{u2}, \ldots, h_{um}, h_{p1}, h_{p2}, \ldots, h_{pk}\}$. We denote this vector as the vertex feature, which will change according to the utterances and personas of different speakers.

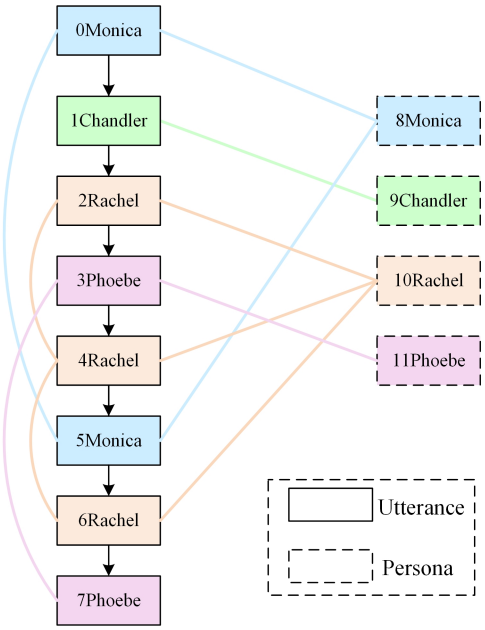

Figure 2: An example of edges construction. Edges of different speakers are marked with different colors. Edges with arrows are one-way sides, and those without arrows have sides in both directions.

Edges: To more fully model the context, we establish three kinds of relationships between the utterances and personas: (i) There is an edge between two adjacent utterances, which points from the one in front of time to the one in back, representing the relationship of time; (ii) There is an edge between the persona of the speaker and all the utterances that belong to the persona of the speaker; (iii) There is an edge between utterances that belong to the same speaker. The constructed edges are shown in Figure 2.

The specific way of building edges is to number all the utterance and persona nodes above starting from 0 and combine the number of the source and target nodes of an edge as the data form of an edge. For example, if utterance node 3 belongs to the persona of the speaker node 11, mark this edge as [3, 11]. In addition, each node is set to have an edge pointing to itself, which is to aggregate the information of neighbor nodes in the graph coding stage without losing the information of the node itself. According to our statistics, each dialogue graph has an average of 10.7 nodes and 24.6 edges.

The edge set $\mathcal{E}$ is initialed as an adjacency matrix and recorded as A, the GCN of each layer is calculated as follows:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}\right) \quad (1)$$

where $\tilde{A} = A + I$ is the adjacency matrix with self connection, I is the unit matrix, $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$, $W^{(l)}$ is the parameter of Layer l, $\sigma(\cdot)$ indicates the activation function, such as $\mathrm{ReLU}(\cdot) = \max(0, \cdot)$. Considering the small scale of the graph constructed by the dialogue, setting the number of layers of GCN to 2 can make each node effectively aggregate the information of adjacent nodes. Such sufficient information can predict the addressee more accurately in the subsequent process. Similarly, the utterance node also aggregates the persona information of the corresponding speaker, which will play an important role as a memory set in the decoding stage.

4.3 Addressee Selecting Mechanism

In multi-party dialogue, selecting the addressee is important to generate an appropriate response. Therefore, this paper incorporates the addressee selecting module. There is no addressee label in the dataset, and the existing methods usually can not supervise the learning of this module. Therefore, the end-to-end method is usually used to update the whole model with the final loss. It is noted that the addressee can be easily inferred from the ground truth response. Inspired by Lian et al. (2019), the ground truth can be regarded as a label to supervise the addressee selecting process in addition to calculating the loss as the standard answer and the generated response. Therefore, this paper adopts the method of a posterior selection and adds a KL divergence to the module as an additional loss during training.

First of all, both during the training stage and the testing stage, the standard process of calculating the addressee is the selection process based on the attention mechanism, which is called prior selection:

$$\begin{aligned} p_{\text{prior}} &= p\left(h_p = h_{pi} \mid h_{p1}\right) \\ &= \frac{\exp\left(h_{pi} \cdot h_{p1}\right)}{\sum_{j=2}^{k} \exp\left(h_{pj} \cdot h_{p1}\right)} \end{aligned} \quad (2)$$

where h_{p1} is the persona representation of the responder, as the query of prior selection attention,

and $h_{pi}(i > 1)$ is the representation of personas of other speakers in this set of conversations, here we adopt dot product attention and softmax for normalization. In the training stage, the ground truth can be used as a label to supervise the training addressee selecting module, that is, the representation of the response can be used as a query to calculate the attention weight with the personas representation of other speakers, which is called posterior selection:

$$\begin{aligned} p_{\text{posterior}} &= p\left(h_p = h_{pi} \mid h_y\right) \\ &= \frac{\exp\left(h_{pi} \cdot h_y\right)}{\sum_{j=2}^{k} \exp\left(h_{pj} \cdot h_y\right)} \end{aligned} \tag{3}$$

where h_y is the representation of the ground truth, as the query of posterior selection attention.

Obviously, the ideal situation is that even if there is no standard answer, the distribution of p_{prior} can be as close as possible to $p_{\text{posterior}}$. Therefore, in addition to negative log likelihood (NLL) loss, this paper also introduces KL divergence as an auxiliary loss other than NLL loss to measure the similarity between p_{prior} and $p_{\text{posterior}}$ .The formula of KL divergence is defined as follows:

$$\text{Loss}_{\text{KL}} = \sum_{i=2}^{k} p\left(h_p = h_{pi} \mid h_y\right) \log \text{K} \tag{4}$$

$$\text{K} = \frac{p\left(h_p = h_{pi} \mid h_y\right)}{p\left(h_p = h_{pi} \mid h_{p1}\right)} \tag{5}$$

In addition, the calculation formula of Loss_{NLL} is:

$$\text{Loss}_{\text{NLL}} = -\sum_{i=1}^{n} \log p\left(y_i \mid y < y_i; X; P\right) \tag{6}$$

where y_i is the word output in the current time step, n is the length of the response. The overall loss of PersonaTKG is:

$$\text{Loss} = \text{Loss}_{\text{NLL}} + \text{Loss}_{\text{KL}} \tag{7}$$

Ultimately, weighted sums of attention weight and personas representation of other speakers are used to obtain the predictive representation of the addressee. The formula is as follows:

$$h_{\text{addr}} = \sum_{i=2}^{k} \alpha_i h_{pi} \tag{8}$$

where i is accumulated from 2, α_i is the value of dimension i in $p_{\text{posterior}}$ (training stage) or p_{prior} (testing stage). Then the personas representation of the responder and the addressee are weighted and summed using the gating mechanism and sent to the decoder. The calculation process is performed as follows:

$$h_p = \alpha_p \cdot W_a h_{p1} + (1 - \alpha_p) \cdot W_b h_{\text{addr}} \tag{9}$$

$$\alpha_p = \sigma\left(V \cdot h_{p1}\right) \tag{10}$$

$$\sigma(x) = \text{Sigmoid}(x) = \frac{1}{1 + \exp(-x)} \tag{11}$$

where W_a, W_b and V are learnable parameters.

4.4 Response Decoder

The decoder is a language model that generates a response word by word based on the context and persona information. Let s_{t-1} be the hidden state of decoder and y_{t-1} be the embedding of word generated in the last time step, h_t is the semantic vector obtained by the weighted sum of attention calculation on the memory set $H = \{h_{u1}, h_{u2}, \ldots, h_{um}, h_{p1}, h_{p2}, \ldots, h_{pk}\}$ of the encoder. Then the output calculation process of GRU in the current time step is:

$$s_t = \text{GRU}\left(\left[y_{t-1}; h_t\right], s_{t-1}\right) \tag{12}$$

The word generated in current time step y_t is obtained after linear transformation and softmax normalization on hidden state s_t.

5 Experiments

5.1 Dataset

The experiments are conducted on our constructed dataset HLA-Chat++. The dataset is divided into train / valid / test set according to the proportion of 96%, 2% and 2%.

5.2 Baselines

We compared PersonaTKG with several strong baselines. To be fair, encoders of all models are implemented with HRED to handle multi-round contexts.

Seq2Seq: a Seq2Seq model with attention mechanism (Sutskever et al., 2014).

DialogueGCN: A dialogue emotion analysis algorithm with a graph neural network encoder (Ghosal et al., 2019). We implement the encoder and add a decoder for generation.

Model	PPL	BLEU-1/2%	Dist-1/2%	Emb E/A/G%	Per R/P/F1%	ACC
Seq2Seq	134.013	9.51/10.57	0.79/2.53	36.05/46.65/41.64	0.02/0.14/0.04	28.3
Per-Seq2Seq	125.889	9.74/10.61	0.97/3.07	36.35/46.74/42.59	0.02/0.12/0.03	28.9
DialogGCN	127.623	9.75/11.03	0.51/1.35	36.89/45.29/42.35	0.02/0.11/0.03	28.6
Per-DialogGCN	125.866	10.55/11.91	0.90/2.95	36.64/46.23/42.23	0.01/0.12/0.02	28.8
SIRNN	119.997	10.32/11.41	0.81/2.64	36.79/48.75/43.19	0.02/0.14/0.04	28.7
Per-SIRNN	120.565	10.75/11.54	0.78/2.43	36.26/48.70/42.67	0.02/0.15/0.05	28.8
PostKS	122.626	10.59/11.37	0.87/2.19	36.68/47.13/42.86	0.02/0.14/0.03	28.9
PersonaTKG+tag	117.063	11.74/12.70	0.85/2.79	36.98/50.09/43.61	0.03/0.18/0.04	29.4
PersonaTKG+doc	114.440	12.09/13.21	1.17/4.25	37.29/51.16/44.13	0.04/0.54/0.05	29.8
PersonaTKG+sent	**109.719**	**13.18/14.17**	**1.59/6.90**	**38.82/53.23/45.83**	**0.05/0.61/0.07**	**30.2**

Table 4: Automatic evaluation results

SIRNN: A multi-party dialogue model with addressee selecting mechanism (Zhang et al., 2018a). We implement the encoder and add a decoder for generation.

Per-: Persona encoder is added to the above three models for persona integration.

PostKS: A persona-based generative network with posterior selection mechanism to guide knowledge selection (Lian et al., 2019). It regards persona as knowledge.

Note that we do not adopt methods such as ALOHA (Li et al., 2020) as baselines, since our method is not built on a pre-trained language model. We leave this potential improvement to the future work.

5.3 Implementation Details

The dimension of word embedding is set to 300 initialized using GloVe (Pennington et al., 2014) pre-trained word vector and the vocabulary size is set to 50000. The word-level, sentence-level encoders in the hierarchical utterance encoder, and persona encoder are single-layer bidirectional GRUs with 800 hidden units and do not share parameters. The GCN is two-layered, and the number of input and output channels is set to 800. In order to facilitate calculation and avoid tedious dimension transformation, the nonlinear layer function after the convolution of the first layer graph is ReLU, and the Dropout Mechanism is used. The batch size is 80 and we use the Adam optimizer with an initial learning rate of 0.0005.

5.4 Automatic Evaluation

We use a variety of automatic evaluation metrics to comprehensively evaluate the performance of PersonaTKG from many aspects.

PPL: perplexity of the model, the smaller PPL score indicates a higher probability of the model producing a real response.

BLEU-1/2: the word-overlap scores of calculating unigrams and bigrams against the ground truth.

Dist-1/2: the proportions of distinct unigrams and bigrams in the generated responses.

Embedding-based metrics: Emb E calculates the semantic similarity between the generated response and the ground truth by averaging word embeddings. **Emb A** and **Emb G** calculate the semantic similarity between the generated response and the ground truth based on average and greedy matching, respectively.

ACC: the accuracy between the ground truth and generated response.

Per R/P/F1: the uni-gram Recall/Precision/F1 scores between the generated response and the persona set (Lian et al., 2019). Specifically, the set of non-stopwords in the generated response is represented by W_Y, and in predefined persona texts are represented by W_P, the calculation formulas of Per R and Per P are as follows:

$$\frac{|W_Y \cap W_P|}{|W_P|} \text{ and } \frac{|W_Y \cap W_P|}{|W_Y|} \tag{13}$$

and Per F1=2 (Per R · Per P) /(Per R + Per P).

Table 4 shows the results of the automatic evaluation, with the best results in bold. On various automatic evaluation metrics, PersonaTKG achieves the optimum compared to the baselines.

It was found that the SIRNN group increases the most, we conjecture that the addressee selecting mechanism of SIRNN plays an important role. By

	Fluency			Persona consistency			Semantic coherence		
	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
PersonaTKG+sent vs PersonaTKG+tag	64	92	44	104	68	28	92	72	36
PersonaTKG+sent vs SIRNN	69	89	42	125	49	26	109	73	18

Table 5: Human evaluation results

Model	PPL	BLEU-1/2%	Dist-1/2%	Emb E/A/G%	Per R/P/F1%	ACC
PersonaTKG+sent	**109.719**	**13.18/14.17**	**1.59/6.90**	**38.82/53.23/45.83**	**0.05/0.61/0.07**	**30.2**
w/o GCN	118.502	10.70/11.86	1.22/3.35	36.75/49.13/43.47	0.02/0.40/0.04	28.9
w/o AS	117.977	11.15/12.53	1.06/2.07	37.15/50.27/43.48	0.02/0.34/0.04	29.1

Table 6: Ablation experiments

predicting the addressee in the encoding stage, a sensitive response to the addressee can be generated according to the given persona information. The improvement of PostKS relative to Seq2Seq proves the effectiveness of a posterior selection mechanism.

DialogGCN and PersonaTKG realize different graph encoding methods, which have a great improvement in BLEU compared with the Seq2Seq model and prove the effectiveness of graph structure encoding. Compared to DialogueGCN, PersonaTKG in the persona related metrics, word embedding metrics, accuracy, and other metrics get higher scores. It shows that considering the construction of various types of nodes in the context, we can extract more sufficient information than simply using the composition of sentence nodes and using the inherent connection between nodes as an edge is a more reasonable choice. The above analysis and experimental results show that PersonaTKG (PersonaTKG + tag) combined with addressee selecting mechanism and specific graph encoding achieves better results.

In addition, after adding two kinds of persona explanations, all metrics are improved compared with only the persona tag model, especially the Per P/R/F1, which show that the introduction of unstructured persona text knowledge can improve the persona consistency in response. However, the effect of adding document-level text knowledge is not good as adding sentence-level text knowledge. We conjecture that document-level text knowledge is too complex and noise is added. We also conducted the t-test on the models' performance. The results show that our PersonaTKG+sent model significantly outperforms the other baseline models with $p<0.01$.

5.5 Human Evaluation

To better evaluate the quality of the generated responses, we performed human evaluation. 200 generated responses were randomly sampled from the test set to 5 graduate students majoring in dialogue system, and compared with two relatively strong baselines in automatic evaluation from the following aspects:

Fluency: The generated response is smooth and free of syntax errors.

Persona consistency: The generated response is consistent with the persona of the speaker.

Semantic coherence: The generated response is semantically coherent to the context.

For each aspect, the evaluator can choose Win, Tie, and Lose. Win outperformed the other, Lose instead, and Tie represents both tied. Based on the results in Table 5, PersonaTKG significantly outperforms other baselines in all subjective evaluation metrics, indicating that PersonaTKG can extract useful information from the context with a wide range of information and complex structure.

5.6 Ablation Experiments

In order to investigate the effect of feature extraction of graph convolution network and the influence of addressee selecting mechanism on response generation, ablation experiments are carried out on these two modules.

w/o GCN: The GCN encoding process was removed.

w/o AS: The addressee selecting mechanism is removed and decoded directly with the persona of the responder.

The results are shown in Table 6. It can be observed that after removing the graph coding and ad-

context	(Doorbell rings. Sydney takes the pail of candy, and answers. Dixon, his wife Diane, and two kids in costumes stand there.) Kids: Trick or treat! Sydney: Hey, guys! There's more candy in there. Come on in! Diane: Good to see you. Sydney: Good to see you. Dixon: Sydney, can I talk to you for a second? Sydney: What's up? Dixon: Sydney, we've been working together a lot of years. I trust you.
Persona	Heroic BSOD They Do Too Dumb To Live Took ALevel In Badass Vitriolic Best Buds Ho Yay (Due to space constraints, only persona tags are given, sentence-level and document-level text knowledge are not given.)
Seq2Seq	i don't know.
Per-Seq2Seq	i don't know.
DialogGCN	i don't know.i don't know.
Per-DialogGCN	i don't know.
SIRNN	Sydney:what's meaning?
Per-SIRNN	Sydney:what?
PostKS	me too.
PersonaTKG+tag	Sydney: i know.
PersonaTKG+doc	Sydney: i know, i trust you, too.
PersonaTKG+sent	Sydney: i know, i trust you, too.

Table 7: Case study.

dressee selecting mechanism, all metrics decrease significantly. After removing the graph coding module, BLEU decreases significantly, indicating that graph coding is helpful for the fluency of generated responses. After removing the addressee selecting mechanism, Distinct decreases significantly, indicating that the addressee selecting mechanism has a great impact on the diversity of generated responses. In addition, we have conducted the t-test on the models' performance. The results show that our PersonaTKG+sent model significantly outperforms the ablated models with $p<0.01$.

5.7 Case Study

Table 7 shows an example of responses generated by different models along with the input message and persona set. It can be seen that simple models are difficult to model complex context, so Seq2Seq and DialogueGCN all generate a general reply *"i don't know."*. SIRNN with addressee selecting mechanism can predict the next speaker *Sydney*, Postks with a posterior mechanism can generate a contextual response *"me too."*.

PersonaTKG can not only predict the addressee but also generate a contextual response. After adding sentence-level and document-level text knowledge, the generated response *"i trust you, too."* is consistent with *Sydney's* persona *Vitriolic Best Buds*. *Sydeny* trust *Dixon* because she regards him as a deep down friend. It is proved that PersonaTKG improves the consistency between responses and personas of the speakers compared with other baselines.

6 Conclusion

In this paper, we first propose to use unstructured text knowledge in MPC to explain the incomprehensive persona tags. We construct a multi-party personalized dialogue dataset HLA-Chat++ based on English drama scripts and propose a model PersonaTKG with addressee selecting mechanism that integrates personas, dialogue utterances, and external text knowledge in a unified graph. The results show that the automatic and human evaluation are superior than other baselines, which demonstrate the effectiveness of our methods in improving persona consistency.

Acknowledgements

We would like to thank the reviewers for their constructive comments. This work was supported by the National Science Foundation of China (61872074, 62106039, 62272092).

References

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 154–164.

Jia-Chen Gu, Hui Liu, Zhen-Hua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021a. Partner matters! an empirical study on fusing personas for personalized response selection in retrieval-based chatbots. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 565–574.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021b. Mpc-bert: A pre-trained language model for multi-party conversation understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 3682–3692.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8342–8360.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. Gsn: A graph-structured network for multi-party dialogues. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5010–5016.

Satwik Kottur, Xiaoyu Wang, and Vitor R Carvalho. 2017. Exploring personalized neural conversational models. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3728–3734.

Ran Le, Wenpeng Hu, Mingyue Shang, Zhenjun You, Lidong Bing, Dongyan Zhao, and Rui Yan. 2019. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1909–1919.

Aaron W Li, Veronica Jiang, Steven Y Feng, Julia Sprague, Wei Zhou, and Jesse Hoey. 2020. Aloha: Artificial learning of human attributes for dialogue agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8155–8163.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 994–1003.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5081–5087.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13343–13352.

Cao Liu, Kang Liu, Shizhu He, Zaiqing Nie, and Jun Zhao. 2019. Incorporating interlocutor-aware context into response generation on multi-party chatbots. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 718–727.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1417–1427.

Bodhisattwa Prasad Majumder, Taylor Berg-Kirkpatrick, Julian J McAuley, and Harsh Jhamtani. 2021. Unsupervised enrichment of persona-grounded dialog with background stories. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*.

Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like hiking? you probably enjoy nature: Persona-grounded dialog with commonsense expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9194–9206.

Zhao Meng, Lili Mou, and Zhi Jin. 2018. Towards neural speaker modeling in multi-party conversation: the task, dataset, and models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8121–8122.

Biswesh Mohapatra, Gaurav Pandey, Danish Contractor, and Sachindra Joshi. 2021. Simulated chats for building dialog systems: Learning to generate conversations from instructions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1190–1203.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical*

Methods in Natural Language Processing (EMNLP), pages 2133–2143.

Zhai Penghua and Zhang Dingyi. 2019. Bidirectional-gru based on attention mechanism for aspect-level sentiment analysis. In *Proceedings of the 2019 11th International Conference on Machine Learning and Computing (ICMLC)*, pages 86–90.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Assigning personality/profile to a chatting machine for coherent conversation generation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4279–4285.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3027–3035.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3776–3784.

Haoyu Song, Yan Wang, Weinan Zhang, Xiaojiang Liu, and Ting Liu. 2020a. Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 5821–5831.

Haoyu Song, Yan Wang, Weinan Zhang, Zhengyu Zhao, Ting Liu, and Xiaojiang Liu. 2020b. Profile consistency identification for open-domain dialogue agents. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6651–6662.

Haoyu Song, Weinan Zhang, Yiming Cui, Dong Wang, and Ting Liu. 2019. Exploiting persona information for diverse generation of conversational responses. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 5190–5196.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4444–4451.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems (NIPS)*, pages 3104–3112.

Ming Tan, Dakuo Wang, Yupeng Gao, Haoyu Wang, Saloni Potdar, Xiaoxiao Guo, Shiyu Chang, and Mo Yu. 2019. Context-aware conversation thread detection in multi-party chat. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6456–6461.

Weishi Wang, Steven CH Hoi, and Shafiq Joty. 2020. Response selection for multi-party conversations with dynamic topic tracking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6581–6591.

Chen Henry Wu, Yinhe Zheng, Xiaoxi Mao, and Minlie Huang. 2021. Transferable persona-grounded dialogues via grounded minimal edits. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2368–2382.

Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. 2018. Hierarchical recurrent attention network for response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5610–5617.

Semih Yavuz, Abhinav Rastogi, Guan-Lin Chao, and Dilek Hakkani-Tur. 2019. Deepcopy: Grounded response generation with hierarchical pointer networks. In *Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue*, pages 122–132.

Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. 2018a. Addressee and response selection in multi-party conversations with speaker interaction rnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, page 5690–5697.

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

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020. A pre-training based personalized dialogue generation model with persona-sparse data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9693–9700.

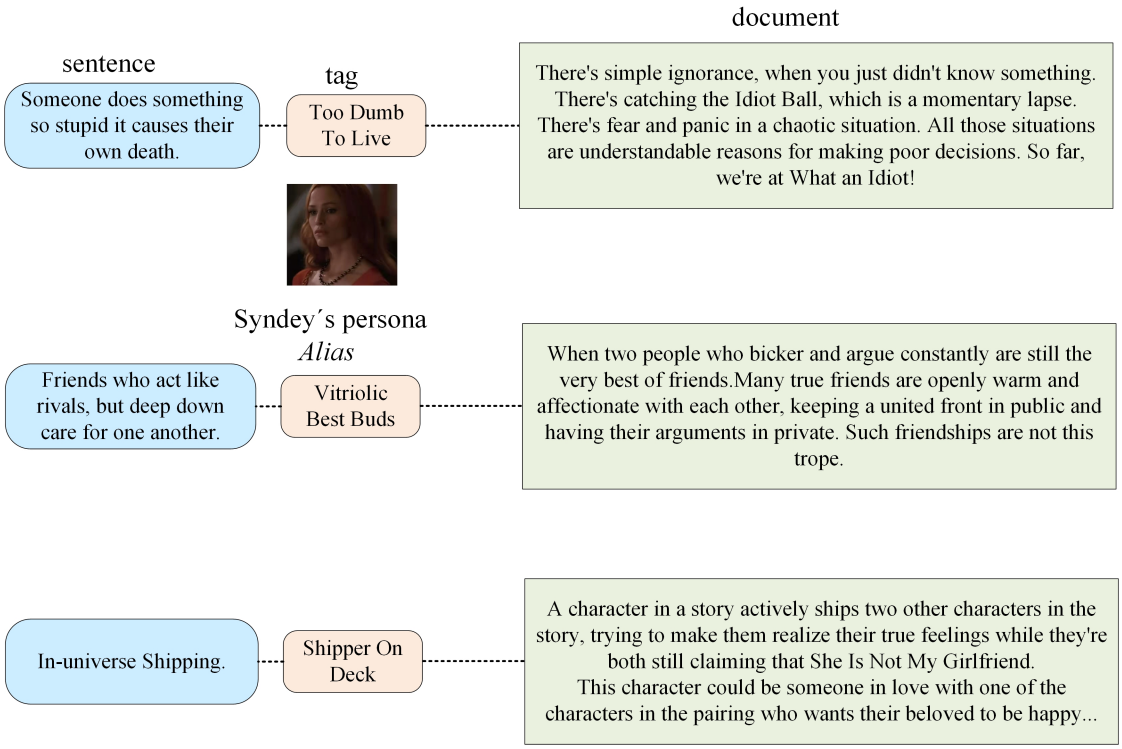

Figure 3: Three personas of Syndey in Alias

A Persona Explanation for Visualization

As shown in in Figure 3, three persona tags *Too Dumb To Live*, *Vitriolic Best Buds* and *Shipper On Deck* of Syndey in *Alias* are given, and explain them on sentence level and document level.