Persona-aware Multi-party Conversation Response Generation

Khyati Mahajan, Samira Shaikh
UNC Charlotte, UNC Charlotte
{kmahaja2, sshaikh2}@uncc.edu

Abstract

Modeling interlocutor information is essential towards modeling multi-party conversations to account for the presence of multiple participants. We investigate the role of including the persona attributes of both the speaker and addressee relevant to each utterance, collected via 3 distinct mock social media experiments. The participants were recruited via MTurk, and were unaware of the persona attributes of the other users they interacted with on the platform. Our main contributions include 1) a multi-party conversation dataset with rich associated metadata (including persona), and 2) a persona-aware heterogeneous graph transformer response generation model. We find that PersonaHeterMPC provides a good baseline towards persona-aware generation for multi-party conversation modeling, generating responses which are relevant and consistent with the interlocutor personas relevant to the conversation.

Keywords: multi-party conversation, dialogue systems, group conversation modeling

1. Introduction

Research in the field of natural language generation (NLG) has often focused mainly on two-party dialogue modeling, with recent advances showcasing capabilities in a hitherto unforeseen manner. A notable challenge has been modeling multi-turn dialogues owing to their non-sequential conversational flows and ambiguity during turn-taking, and modeling speaker information such as persona attributes, towards more consistent and relevant response generation (Ni et al., 2023; Zhang et al., 2018c; Li et al., 2016).

Conversations with more than 2 participants, or multi-party conversations (MPCs), are just as prevalent in everyday life, and research in modeling MPCs has seen a recent rise. The presence of multiple participants poses new and interesting challenges for MPC modeling. While the need to account for initiative taking is similar to multi-turn two-party modeling, MPCs require accounting for speakers and addressees for each turn. Most recent research has focused on modeling 1) speaker and addressee information (Qiu et al., 2020), 2) response selection (a retrieval task) (Lowe et al., 2015; Wu et al., 2017; Zhou et al., 2018; Tao et al., 2019; Gu et al., 2020; Jia et al., 2020; Wang et al., 2020) or response generation (a generation task) (Zhang et al., 2018a; Liu et al., 2019; Hu et al., 2019; Gu et al., 2022) with some papers 3) jointly modeling both (Ouchi and Tsuboi, 2016; Zhang et al., 2018b; Le et al., 2019; Zhang et al., 2018b; Gu et al., 2021). It has been shown that modeling interlocutor (or user, used interchangeably in the paper) information often outperforms standalone response selection or generation tasks. However, the investigation of how attributes related to the users - such as persona - might affect the response generation capabilities for MPCs has been limited (Ju et al., 2022). Thus, we aim to study the effect of including speaker and addressee personas towards response generation for multi-party conversation modeling.

Specifically, we study how user attributes such as race, gender, leaning, and behavior type on social media might contribute towards generating responses that are more relevant and consistent in keeping with the involved interlocutors We present PersonaHeterMPC, based on HeterMPC (Gu et al., 2022) towards this task (Section 4). We follow automatic evaluation strategies similar to HeterMPC (Gu et al., 2022), adding human evaluations not just for checking 1) relevance, 2) fluency and 3) informativeness of the response (similar to HeterMPC) but also appointing scores for 4) initiative-taking (to check whether the response helps move the conversation along), 5) thread response appropriateness (to check whether the response is relevant for the thread within the conversation), and 6) persona-relevancy (whether the response is relevant according to the speaker and addressee personas). Our main contributions include 1) a persona-aware multi-party conversation dataset and 2) a persona-aware response generation model which utilizes heterogeneous graph transformers.

2. Related Work

We begin with related work towards response generation in MPC modeling, then focus on persona-level datasets and existing work in persona MPC modeling. We limit discussion to research focused solely on MPC modeling since it is more central to our goal than the substantial work on persona related two-party dialogue modeling.

Response Generation. Zhang et al. (2018a) propose a tree-based model frame for structure-aware group conversations, organizing the group

LREC-COLING 2024, pages 12712–12723
20-25 May, 2024.

conversation as a tree with different branches involving multiple conversation threads. They utilize hierarchical encodings with the Seq2Seq encoder-decoder model (Sutskever et al., 2014) implemented with GRUs (Chung et al., 2014). They outperform approaches evaluated on two-party dialogue modeling with the Ubuntu Corpus (Lowe et al., 2015). Liu et al. (2019) propose incorporating Interlocutor-aware Contexts into Recurrent Encoder-Decoder frameworks (ICRED), leveraging an addressee memory mechanism to enhance contextual interlocutor information for the addressee, predicting both speaker and addressee when generating responses. Comparison of ICRED with other research is difficult owing to evaluation on differing datasets, but the authors find that it outperforms two-party dialogue models Seq2Seq, PersonaModel (Li et al., 2016), and VHRED (Serban et al., 2017) on their dataset. Hu et al. (2019) generalize existing sequence-based models to a Graph-Structured neural Network (GSN) for dialogue modeling, using a graph-based encoder that can model the information flow. They utilize the Ubuntu Corpus and find that GSN outperforms Seq2Seq and HRED (Serban et al., 2016) (succeeded by VHRED (Serban et al., 2017)), both trained towards two-party dialogue modeling. Recently, (Gu et al., 2022) present HeterMPC, a heterogeneous graph transformer for MPC response generation, with 2 types of nodes representing utterances and interlocutors, and 6 meta-relations node-edge-type-dependent parameters to characterize the heterogeneous interactions. They evaluate over the Ubuntu Corpus outperforming Seq2Seq (Sutskever et al., 2014), Transformers (Vaswani et al., 2017), and GSN by a statistically significant margin. We base our model architecture on HeterMPC owing to its performance compared to previously proposed approaches to model persona-level attributes (Section 4).

Modeling persona attributes. Persona related research in MPC modeling is limited, with PersonaTKG (Ju et al., 2022) being the only proposed model to the best of our knowledge. They utilize hierarchical encoding, with the utterance encoder consisting of word-level and sentence-level encoders with bidirectional GRUs. They model utterance and persona nodes in a homogeneous manner, with the dialogues concatenated to represent a vertex in the graph. The edges model 3 relationships, between 1) an utterance and its reply (and vice versa), 2) between the persona of the speaker and all the utterances that belong to the persona of the speaker, and 3) between utterances that belong to the same speaker. The model is evaluated on HLA-Chat++[1], a dataset created by the authors, and compared with Seq2Seq, DialogueGCN (Ghosal et al., 2019), SIRNN (Zhang et al., 2018b), and PostKS (Lian et al., 2019), outperforming the models (which are modified to include persona representations to allow comparisons).

It is important to note that PersonaTKG follows a different modeling approach than our main aim. They utilize Graph Convolutional Networks (GCNs), whereas HeterMPC (and thus our model) utilize Transformers and (heterogeneous cross) attention, which have been shown to be more effective for modeling textual information. A closer look at the dataset they utilize (HLA-Chat++) also reveals that extracting relevant fields towards modeling is not straightforward (refer to Table 4), and scripts for performing this are not provided, making it difficult to utilize the dataset towards our task. Furthermore, while HLA-Chat++ is similar to our dataset in terms of informal conversations, it is a scripted dataset, whereas our study requires a real-world unscripted dataset for open domain conversation modeling.

	Exp 1	**Exp 2**	**Exp 3**
Time	Apr 2021	Oct 2021	Mar-Apr 2022
Race	80% W, 20% M	77% W, 23% M	81% W, 19% M
Gender	50% F, 49% M, 1% O	57% F, 42% M, 1% O	52% F, 47% M, 1% O
Leaning	51.5% L, 42.5% C, 6% I	42% L, 41% C, 17% I	51% L, 44% C, 5% I

Table 1: Data collection statistics - Race is white (W), minority (M); Gender is female (F), male (M), other (O); Leaning is conservative (C), liberal (L), Independent (I). Categories have been crudely simplified for modeling.

Our search for MPC with persona-level attributes thus continues with a recent survey on this topic (Mahajan and Shaikh, 2021), which lists two relevant corpora. The FriendsPersona corpus (Jiang et al., 2020) does not provide user level personas and the TEAMS entrainment corpus (Litman et al., 2016) does not provide the explicit speakers and addressees of each utterance in the conversation. Another corpus which could be useful is the PersonaChat corpus (Zhang et al., 2018c), however this also does not have conversation level data with defined speakers and addressees, and corresponding personas. Our modeling task involves utterance-level speakers and addressees, along with their personas (which are a constant property of the user). This property is not available for these datasets - another reason we collect and create our dataset (Section 3). We also considered MultiLIGHT (Wei et al., 2023), consisting of fantasy-based triadic conversations, but this meant limiting the modeling to triadic conversations, and limited personas which were not reflective of online personas. Lastly, there was the possibility of synthetically generating conversations as presented in PLACES (Chen et al., 2023), and we consider this method future work to bolster datasets.

Evaluation. Evaluation strategies for NLG have

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

been a hot and debated topic for a long time (Howcroft et al., 2020; Agarwal et al., 2020). The focus has remained on two-party dialogue modeling generations, with benchmarks proposed towards improving comparisons across research to place progress better (Gehrmann et al., 2021; Liu et al., 2021). However, research on this front quite often does not include multi-party response generation, and although the difference in generating utterances is not be very different, (Mahajan et al., 2022) point to the shortcomings in existing MPC modeling research when it comes comparing performance across work. We ensure to utilize the evaluation methods proposed in HeterMPC for consistency in reporting (Section 5), and report additional human evaluation metrics towards multi-party specific challenges.

Exp No.	Total Users	Annotated by	Behaviors			
			Avoiders	Expressors	Spectators	Suppressors
1	121	`flan-t5-xxl`	7	62	41	11
		Manually corrected	18	76	19	8
2	140	`flan-t5-xxl`	7	70	46	17
		Manually corrected	12	91	23	14
3	182	`flan-t5-xxl`	10	102	66	4
		Manually corrected	23	111	38	10

Table 2: User behavior annotation statistics

3. Dataset

For the purposes of our experiment, we require a controlled environment where we can ask for explicit consent to collect data, and collect persona-level attributes of the participants and connect the personas with their social media posts in differing environments. These experimental conditions are difficult to collect via existing social media platforms, and thus we utilize a mock social media platform (Mahajan et al., 2021) (Section 3.1). We derive automatic annotations based on the behaviors exhibited by the users on the platform, which add to the users' persona behaviors (Section 3.2). We conclude with Section 3.3.

3.1. User Experiment Setup

We simulate a mock social media network environment for collecting data to enable the observation of users in differing environments. We collect data over 3 distinct experiments for this IRB approved study to ensure diversity in the topics being discussed. We follow guidelines listed in (Mahajan and Shaikh, 2021) towards dataset creation, and make sure to remove personally identifiable information (PII) before utilizing the dataset. Much of our data collection efforts were underway during the COVID-19 pandemic, and our efforts towards equitable distributions in our participant pool were difficult (Table 1). Moreover, many conversations on the platform tended to focus around this topic. A larger team comprising of interdisciplinary researchers was involved to ensure participation on the platform that reflected the behaviors observed in emotional firestorms. The research team collected informed consent from all participants, which included details of how the data could be utilized in related research.

Utilizing Community Connect (Mahajan et al., 2021), we construct a structured social network with roughly 15 sub-groups within the network. Each group is designed such that it is either heterogeneous (good mix of liberal and conservative leaning users around 50-50) or homogeneous (overall liberal or conservative leaning around 80-20). The social network is connected via bridge users, which connect groups in differing ways (e.g. connecting a heterogeneous group to a homogeneous liberal leaning group). For the scope of this paper, qualitative findings from the data collection and user behaviors are considered future work.

```
Instruction: Classify User1 into one of the 4 categories
as defined below:
Spectators: <definition>
Expressors: <definition>
Avoiders: <definition>
Suppressors: <definition>
User1: I still don't think we should have to have proof
of vaccinations to go anywhere, when masks were supposed
to be working all along.
User2: You already need proof of several other vaccina-
tions in order to attend school, go abroad, or work in
certain fields.
User1: That is true but this is slightly different, it's
too new for some
```

Figure 1: Zero-shot prompt example to generate annotations for each user using `flan-t5-xxl`

3.2. Social Media Behavior Categories and Annotation Methodology

One of the motivations for our study is to study how providing persona inputs can generate responses tailored to a specific behavior for participating in the MPC. We focus on 4 main categories of behaviors that are observed during an emotional firestorm - *Spectators, Expressors, Avoiders,* and *Suppressors* (Gross, 1998).

Spectators are defined as participants who prefer to observe emotional conversations unfolding, and utilize social media as a place to obtain information from or a place to keep in contact with family and friends not share firestorm content. *Expressors* tend to utilize social media to seek, process, and express emotions. They find the spread of emotions to be a positive goal in and of itself, and often can be seen to spread content based upon its connotation of being "powerful" or because it "needs to be heard." They are much less likely to consider social media any different a place to engage in firestorm content than a real world conversation. *Avoiders* are discerning and cautious in their emotion sharing on social media, preferring to discuss difficult topics but mainly sharing content they find positive, unifying, or productive. *Suppressors* suppress overly emotional content on social media during a firestorm, viewing intense emotion expression on social media (and hence Expressors' posts) as orthogonal to productive discourse. Critically, instead of avoiding emotional social media content like the Avoiders, they actively engage in discourse with Expressors by attempting to advance facts and advocate for suppressing the emotion expression.

We utilize `flan-t5-xxl` (Chung et al., 2022) owing to its performance towards similar classification tasks (Chia et al., 2023), to ease the computational burden on our annotators and reduce the time required to gather annotations. We use zero-shot prompts to generate annotations reflecting the typical behavior of each user (Figure 1). We experiment with prompt variations, most notably trying to generate annotations for all users at once, but find that the model performs more deterministically when users are explicitly mentioned.

Statistic	Exp 1	Exp 2	Exp 3
Conversations > 5 utterances	550	563	720
Total no of turns	6384	5242	9845
Avg turns per conversation	11.61	9.31	13.67
Total no of tokens	97142	83995	144615
Avg tokens per turn	15.22	16.02	14.69
Avg tokens per conversation	15.71	15.55	14.34
Vocab size	9039	8520	11047
Total users	122	144	187
Avg users in conversation	6.55	6.75	9.41

Table 3: Final dataset statistics

Once the annotations are generated for each user, they are manually checked for accuracy by 2 graduate student annotators. On average, they find 70.1% annotations reflect the user behavior well, whereas 29.9% annotations are modified to reflect user behavior better. The statistics for the annotations for each category are provided in Table 2. When asked whether the `flan-t5-xxl` were helpful, the annotators claimed that the automatic annotations provided a good baseline which made their task easier and faster. It is worth noting that Expressors form a clear majority of behaviors in our experiments, whereas Suppressors are fewer in number. Most users classified as Avoiders did not post much during the entire experiment, whereas those classified as Spectators preferred engaging with non-political content.

3.3. Dataset Discussion

The data from each experiment is collected into a common dataset, of which 20% (521 conversations) is randomly sampled as test data, 15% (313 conversations) is randomly sampled as validation data, and the remaining 65% (1766 conversations) is used as training data (more statistics in Table 3). The dataset is available upon request[2], with access contingent upon approval. Refer to Table 4 for details of how we construct the final data utilizing Community Connect fields.

Input field for modeling	Description	Field from Community Connect
context	All utterances in the conversation, other than the utterance to be generated	Text from body of posts, except the one to be generated
relation_at	List of lists which describes how each context utterance is related to its parent in the conversation graph	Determined by computing the utterance_turn based on parent_id and feed_id
ctx_spk	Relative IDs of the speakers of the utterances in the context	Determined by taking the user_handle of all the speakers in the context, and computing their user ID for the conversation
ctx_adr	Relative IDs of the addressees of the utterances in the context	Determined by taking the user_handle of all the addressees in the context, and computing their user ID for the conversation
answer	Utterance to be generated	Text from body of posts to be generated
ans_idx	The position of the node where the response will be added into the graph	Determined by computing the utterance_turn based on parent_id of the utterance to be generated
ans_spk	Relative ID of the speaker of the utterance to be generated	Determined by taking the user_handle of the speaker of the utterance to be generated, and computing their user ID for the conversation
ans_adr	Relative ID of the addressee of the utterance to be generated	Determined by taking the user_handle of the addressee of the utterance to be generated, and computing their user ID for the conversation
ctx_spk_persona	List of personas for each speaker in context	Compiled from user survey
ctx_adr_persona	List of personas for each addressee in context	Compiled from user survey
ans_spk_persona	Speaker persona of utterance to be generated	Compiled from user survey
ans_adr_persona	Addressee persona of utterance to be generated	Compiled from user survey

Table 4: Input fields required for modeling, and how they were derived from the dataset collected via Community Connect

4. Response Generation Model

Owing to the capabilities of HeterMPC[3] (Gu et al., 2022) in 1) modeling the MPC as a heterogeneous conversation graph (Hu et al., 2019), 2) response generation of an utterance anywhere in the graph, and 3) support for utilizing the attention mechanism and Transformers for modeling (Hu et al., 2020), we base our model on it (Section 4.1). We utilize HeterMPC given its performance towards modeling the Ubuntu corpus (Lowe et al., 2015) - which has similar properties in terms of informal, asynchronous conversations. We approach multiple persona modeling techniques, and discuss the implications (Section 4.2).

4.1. Background

A heterogeneous graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ is used to model the relationships between $\mathcal{V}$ nodes (which are utterance $\mathcal{M}$ or interlocutor $\mathcal{I}$ type) with $\mathcal{E}$ edges. $\mathcal{E} = \{e_{p,q}\}_{p,q=1}^{M+I}$ is the set of directed edges, between nodes p and q. Six types of meta relations $\{reply, replied\text{-}by, speak, spoken\text{-}by, address, addressed\text{-}by\}$ describe the directed edge between two graph nodes (Sun et al., 2011, 2013). If an utterance represented by node n replies another utterance represented by node m,

[2]https://forms.gle/NCgc62aYUrb9SGuX8

[3]https://github.com/lxchtan/HeterMPC

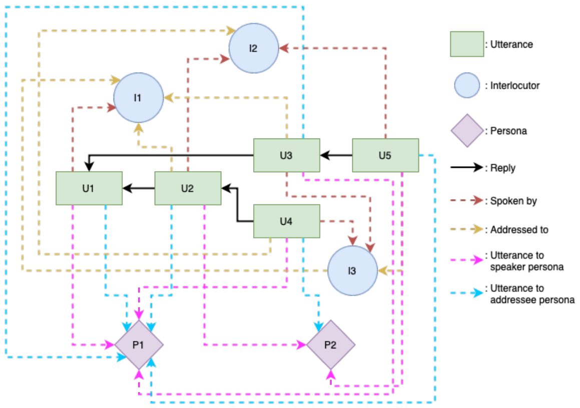

Figure 2: Example conversation graph. All edges have bi-directional counterparts.

the edge $e_{n,m} = reply$ and the reversed edge $e_{m,n} = replied\text{-}by$. If an utterance represented by node m is spoken by an interlocutor represented by node i, $e_{i,m} = speak$ and $e_{m,i} = spoken\text{-}by$. If an utterance represented by node n addresses an interlocutor represented by node i, $e_{n,i} = address$ and $e_{i,n} = addressed\text{-}by$. In other cases, $e_{p,q} = NULL$ to indicate no connection between nodes p and q. These are showcased in Figure 2 with one-way connections for brevity.

Node initialization. Each node in HeterMPC is represented as a vector, with utterances encoded by a `[CLS]` token inserted at the start of each utterance, and a `[SEP]` token inserted at the end (Devlin et al., 2019). The Transformer architecture is utilized to encode and learn contextual representations (Vaswani et al., 2017). The calculation for an utterance at the l-th Transformer layer is denoted as $\mathbf{H}_m^{l+1} = \textit{TransformerEncoder}(\mathbf{H}_m^l)$, where $m \in \{1, ..., \mathcal{M}\}$ and $l \in \{0, ..., L_1 - 1\}$, L_1 denotes the Transformer layers for initialization, $\mathbf{H}_m^l \in \mathcal{R}^{k_m \times d}$, k_m denotes the length of an utterance and d denotes the dimension of embedding vectors. Interlocutors nodes are directly represented with an embedding vector, derived by looking up an order-based interlocutor embedding table (Gu et al., 2020). Since the order of each interlocutor is determined relative to their utterance in a given conversation, it can be used across train, validation, and test sets.

Heterogeneous Attention. If (s, e, t) denotes an edge e connecting a source node s to a target node t, l-th iteration representations denoted by $\mathbf{h}_s^l$ and $\mathbf{h}_t^l$. The heterogeneous attention weight $w^l(s, e, t)$ before normalization is calculated as:

$$\mathbf{k}^l(s) = \mathbf{h}_s^l \mathbf{W}_{\tau(s)}^K + \mathbf{b}_{\tau(s)}^K, \tag{1}$$

$$\mathbf{q}^l(s) = \mathbf{h}_s^l \mathbf{W}_{\tau(t)}^Q + \mathbf{b}_{\tau(t)}^Q, \tag{2}$$

$$w^l(s, e, t) = \mathbf{k}(s) \mathbf{W}_{e_{s,t}}^{ATT} \mathbf{q}(t) \frac{\mu_{e_{s,t}}}{\sqrt{d}}, \tag{3}$$

where $\tau(s), \tau(t) \in \{UTR, ITR\}$ distinguish utterance (UTR) and interlocutor (ITR) nodes. Eqs. 1 and 2 are node-type-dependent linear transformations. Eq. 3 contains an edge-type-dependent linear projection $\mathbf{W}_{e_{s,t}}^{ATT}$ where $\mu_{e_{s,t}}$ is an adaptive factor scaling to attention. All $\mathbf{W}^* \in \mathcal{R}^{d \times d}$ and $\mathbf{b}^* \in \mathcal{R}^d$ are parameters to be learnt.

Heterogeneous Message Passing. When passing the message of a source node that serves as a value (V) vector to a target node, node-edge-type-dependent parameters are also introduced considering the heterogeneous properties of nodes and edges. Mathematically:

$$\bar{\mathbf{v}}^l(s) = \left(\mathbf{h}_s^l \mathbf{W}_{\tau(s)}^V + \mathbf{b}_{\tau(s)}^V\right) \mathbf{W}_{e_{s,t}}^{MSG}, \tag{4}$$

where $\bar{\mathbf{v}}^l(s)$ is the passed message and all $\mathbf{W}^* \in \mathcal{R}^{d \times d}$ and $\mathbf{b}^* \in \mathcal{R}^d$ are parameters to be learnt.

Heterogeneous Aggregation. All source node messages need to be aggregated for the target node:

$$\bar{\mathbf{h}}_t^l = \sum_{s \in S(t)} softmax(w^l(s, e, t)) \bar{\mathbf{v}}^l(s), \tag{5}$$

where $S(t)$ denotes the set of source nodes. The summarized message $\bar{\mathbf{h}}_t^l$ is aggregated with the original node representation $\mathbf{h}_t^l$ (He et al., 2016) as:

$$\mathbf{h}_t^{l+1} = FFN_{\tau(t)}(\bar{\mathbf{h}}_t^l) + \mathbf{h}_t^l \tag{6}$$

When stacking L_2 iterations, a node can attend to other nodes up to L_2 hops away. The utterance node update at the l-th iteration is then compressed by a linear transformation as:

$$\hat{\mathbf{h}}_t^{l+1} = [\mathbf{h}_t^l; \mathbf{h}_t^{l+1}] \mathbf{W}_{com} + \mathbf{b}_{com}, \tag{7}$$

where $\mathbf{W}_{com} \in \mathcal{R}^{2d \times d}$ and $\mathbf{b}_{com} \in \mathcal{R}^d$ are parameters to be learnt. $\hat{\mathbf{h}}_t^{l+1}$ replaces the utterance representation of [`CLS`] (i.e., $\mathbf{h}_t^l$) in the sequence representations of the whole utterance. Finally, the updated sequence representations are fed into the additional Transformer layer for another round of intra-utterance self-attention, so that the context information learnt by the [`CLS`] representation can be shared with other tokens in the utterance.

Decoder. The standard Transformer model is utilized to generate responses (Figure 4). The cross-attention operation over the node representations of the graph encoder output is performed to incorporate graph information for decoding, followed by a residual connection along with layer normalization. The representations for the response to be generated are masked during training. L_3 denotes the number of decoder layers.

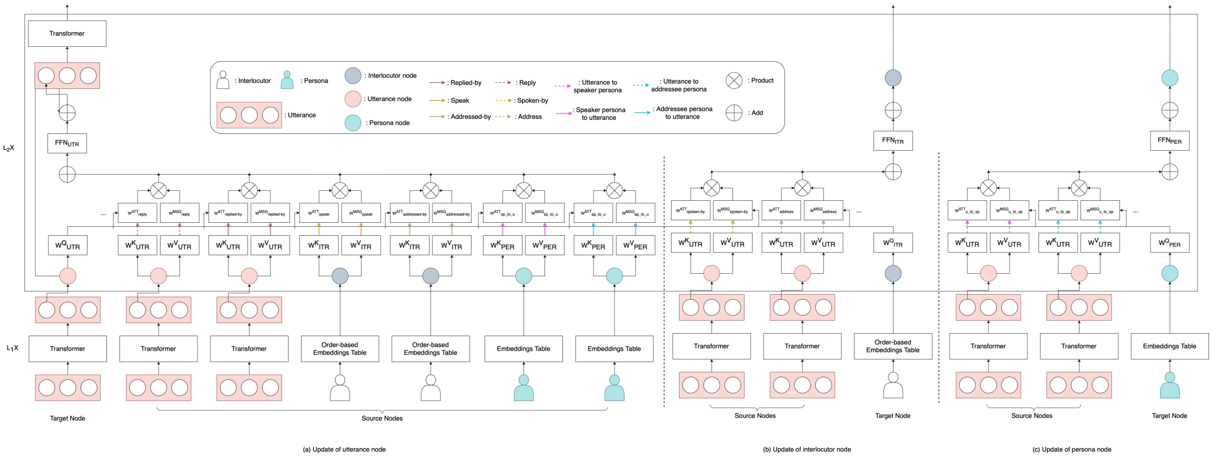

Figure 3: PersonaHeterMPC, based on HeterMPC (Gu et al., 2022). Model details in Section 4.2. The colors for the graph relations are coded similar to the relations showcased in Figure 2.

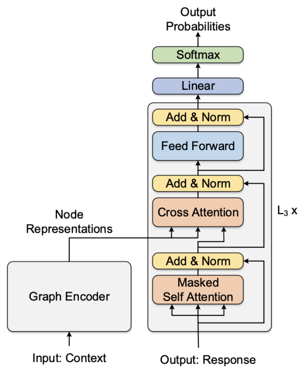

Figure 4: HeterMPC (Gu et al., 2022) Decoder

4.2. Model Architecture

We study two approaches towards persona-aware generation. Approach 1 involves modeling personas concatenated with utterance encodings, and Approach 2 involves modeling personas as new node types and adding edges to connect them to utterances (Figure 3).

PersonaHeterMPC$_{\textbf{concat}}$. The input encodings consist of the speaker persona, addressee persona, and utterance encoding. The input thus changes from being the encoded context $H = \{h_{u1}, ...\}$ to also including the persona attributes $H = \{(p_{u1}spk, p_{u1}adr, h_{u1}), ...\}$. Inputs to the decoder for generation consist of the a concatenated vector which includes the speaker persona, addressee persona, and the [BOS] token $D = \{p_{ans}spk, p_{ans}adr, \texttt{[BOS]}\}$. The computation for loss is updated to reflect the persona inputs by marking their positions with tokens indices set to $[-100]$, thus not including the inputs towards calculating the performance of response generation.

PersonaHeterMPC$_{\textbf{graph}}$. We create persona graph nodes and edges that model the relationships of speaker and addressee personas to an utterance. For heterogeneous graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, $\mathcal{V}$ becomes a set of $M + I + K$ nodes. We introduce four new meta-relations for persona connections (in addition to the six existing edge types), namely $\{utt\text{-}to\text{-}spk\text{-}persona, spk\text{-}persona\text{-}to\text{-}utt, utt\text{-}to\text{-}adr\text{-}persona, adr\text{-}persona\text{-}to\text{-}utt\}$ along with the six meta-relations for utterance and interlocutor edges. If an utterance represented by node n is spoken by an interlocutor whose persona is represented by node s, $e_{n,s} = utt\text{-}to\text{-}spk\text{-}persona$ and $e_{s,n} = spk\text{-}persona\text{-}to\text{-}utt$. If an utterance represented by node n is spoken to an interlocutor whose persona is represented by node a, $e_{n,a} = utt\text{-}to\text{-}adr\text{-}persona$ and $e_{a,n} = adr\text{-}persona\text{-}to\text{-}utt$. Since our aim is to study how different speaker and addressee persona properties affect response generation, the persona nodes are initialized by indexing globally over the entire dataset with a global lookup table, and modeled with an embedding vector calculated on the basis of this value.

5. Experiments and Results

To support comparisons in future work, we follow the evaluation strategies detailed in HeterMPC (Gu et al., 2022). Similar to previous work (Hu et al., 2019), we utilize the evaluation package released by Chen et al. (2015) for BLEU-1 to BLEU-4, METEOR and ROUGE$_L$. We also perform human evaluation to measure 1) relevance, 2) fluency and 3) informativeness, along with 4) initiative-taking to check whether the response helps move the conversation along, 5) thread response appropriateness to check whether the response is relevant for the thread within the conversation, and 6) persona-relevancy whether the response is relevant according to the speaker and addressee personas.

5.1. Response Generation Experiments

Much of the training hyperparameters were set similar to those of HeterMPC$_{\text{BERT}}$, utilizing `bert-base-uncased` pre-trained weights (Wolf et al., 2020), optimization with AdamW (Loshchilov and Hutter, 2017), max gradient norm 1.0, layers for initializing utterance representations (L_1) 9, layers for heterogeneous graph iteration (L_2) 3, and number of decoder layers (L_3) 6. The maximum utterance length was 50, and the max persona length was set to match this at 50. We also changed the batch size to 4, and the gradient accumulation steps to 2 (owing to the dataset size). The validation set was used to select the best model for testing. The decoding strategy was changed to sampling instead of greedy decoding, and we experiment with different top_p and top_k values. All experiments were run on a single A100 GPU. The maximum number of epochs was set to 30, taking about 8 hours. We release our code to allow reproduction of our results. We experiment with HeterMPC$_{\text{BERT}}$ (hereto referred to as HeterMPC in this work) since our dataset size is much smaller than the Ubuntu Corpus, and the suitability of BERT training towards our task. We also tried various learning rates, but found that 6.25×10^{-5} performed best. We aim to experiment with HeterMPC$_{\text{BART}}$ in future work.

We experiment with laconic vs descriptive persona attributes, and find that the descriptive personas perform better. Descriptive personas are generated using a template. For example, if the persona attributes of a person state "`white female liberal expressor`", the descriptive persona would translate to "'*I am a white female with a liberal ideology. I usually prioritize emotional expression on social media, and view it as a platform to share powerful and important content.*"

5.2. Evaluation

To support comparisons in future work, we follow the evaluation strategies detailed in HeterMPC (Gu et al., 2022). Similar to previous work (Hu et al., 2019), we utilize the COCO evaluation package (Chen et al., 2015) for BLEU-1 to BLEU-4, METEOR and ROUGE$_{\text{L}}$. We also perform human evaluation to measure 1) relevance, 2) fluency and 3) informativeness, along with 4) initiative-taking to check whether the response helps move the conversation along (based on subjective measures - mainly recovery and cooperativity - as discussed in (Allen et al., 1999)), 5) thread response appropriateness to check whether the response is relevant for the thread within the conversation, and 6) persona-relevancy whether the response is relevant according to the speaker and addressee personas.

We present the results for three main response generation experiments in Table 5 - (1) the original HeterMPC model without persona information, (2) persona information modeled along with utterance encodings (PersonaHeterMPC$_{\text{concat}}$), and (3) persona information modeled as graph nodes with edges connected to utterance nodes (PersonaHeterMPC$_{\text{graph}}$).

We find that PersonaHeterMPC$_{\text{graph}}$ performs better in automatic evaluations. We also utilize a few other combinations for hyperparameters, notably $(top\_p = 0.3, top\_k = 10)$ which performs very well on automatic evaluations for HeterMPC. However, we find that many generations in these hyperparameters are NaNs (around 14%). In comparison, most generations for the hyperparameters we report in Table 5 produce fewer NaNs (about 5% to 7%). Thus, we include the generations obtained from these hyperparameters combinations. Additionally, we recognize that generations might be affected by responses being images or gifs instead of text, and thus multimodal modeling for multi-party conversations is part of future work.

We report the average ratings given by two expert annotators in Table 6. We find that PersonaHeterMPC$_{\text{graph}}$ performs comparable to HeterMPC on utterance-level measures (relevance, fluency, informativeness) and better on conversation-level measures (initiative-taking, thread relevance, persona relevance). We also calculate Cohen's κ for interrator agreement, and find that most scores are either weak or chance agreement. However, this agreement is also reflected for human ground truth evaluations. This points to the possibility that the annotation task is highly subjective, and thus we report the average scores. Along with the automatic metrics, we hope the average scores can provide some insight into how the models perform towards the persona-aware MPC response generation task. To further investigate the performance, we conduct case studies on all models and study the outputs generated manually.

We vary the speaker persona for case studies (one example is included in Table 7. We find issues with fluency especially for PersonaHeterMPC$_{\text{concat}}$ - both HeterMPC and PersonaHeterMPC$_{\text{graph}}$ perform better. PersonaHeterMPC$_{\text{graph}}$ responses are more in keeping with the political and thus emotional charge of the conversation as well as the speaker persona.

6. Conclusion

We contribute an MPC dataset with persona attributes for each speaker and addressee on an utterance level. We obtain persona information 1) via surveys during participant recruitment in the mock social media experiments, and 2) by annotating observed behaviors based on participant interaction on the platform. We find that zero-shot prompt-

Model	(top_p, top_k)	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE_L
HMPC	(0.9, 5)	12.091	4.967	2.558	1.701	5.076	**9.377**
PHMPC_c	(0.9, 5)	**13.118**	4.740	2.066	1.121	4.960	6.979
PHMPC_g	(0.9, 5)	12.784	**5.834**	**3.697**	**2.859**	**5.338**	9.013
HMPC	(0.5, 5)	11.712	4.894	2.940	2.244	4.978	**9.612**
PHMPC_c	(0.5, 5)	11.305	4.358	2.068	1.285	4.594	6.574
PHMPC_g	(0.5, 5)	**12.367**	**5.643**	**3.652**	**2.902**	**5.153**	9.020
HMPC	(0.9, 10)	11.747	4.696	2.727	1.993	4.869	**8.263**
PHMPC_c	(0.9, 10)	**12.085**	4.125	1.293	0.468	4.452	6.420
PHMPC_g	(0.9, 10)	11.856	**5.009**	**2.861**	**2.036**	**5.052**	8.244
HMPC	(0.5, 10)	11.396	4.788	2.842	2.126	4.856	**9.460**
PHMPC_c	(0.5, 10)	10.533	3.809	1.616	0.961	4.509	6.678
PHMPC_g	(0.5, 10)	**12.473**	**5.566**	**3.510**	**2.725**	**5.120**	8.733

Table 5: Automatic evaluations for PersonaHeterMPC - concat (PHMPC_c) and graph (PHMPC_g) compared to HeterMPC (HMPC) with different generation hyperparameters (top_p, top_k) - best values are in bold.

Models	Max	Human	HMPC	PHMPC_{concat}	PHMPC_{graph}
Relevance	1	0.766	0.266	0.133	**0.433**
Fluency	1	0.966	**0.566**	0.233	0.466
Informativeness	1	0.8	**0.166**	0.033	0.000
$\text{Utterance-level}_{avg}$	3	2.533	**1.000**	0.400	0.900
Initiative-taking	1	0.700	**0.166**	0.000	0.100
Thread relevance	1	0.733	0.233	0.133	**0.366**
Persona relevance	1	0.733	0.366	0.266	**0.466**
$\text{Conversation-level}_{avg}$	3	1.466	0.600	0.400	**0.833**

Table 6: Human evaluation scores (averaged) for evaluating ground truth (Human), HeterMPC (HMPC) and PersonaHeterMPC (PHMPC) with utterance encodings and graph based modeling.

ing on instruction trained `flan-t5-xxl` provides a great behavior annotation baseline, with manual checks showing around 70% accuracy for labeling. We then study the performance of a response generation model, focusing on whether providing personas as inputs leads to an improvement in performance. We find that including persona attributes as graph nodes improves over HeterMPC trained without persona attributes.

One area of future work revolves around the dataset size, which is quite small compared to the Ubuntu corpus ($\tilde{5}0x$ smaller). Similar to social media, some posts contain images, gifs, and emojis instead of text, pointing to future work with multi-modal modeling. Owing to resource and time constraints, studies with network structure changes (L_1, L_2, L_3) also remain next steps. Another area for future work is to bolster our current dataset by synthetic MPC generation (Chen et al., 2023).

Limitations

A major limitation of modeling personas in MPC is the lack of resources which contain all the information required for the task. This limitation affects our work, as the results can only be evaluated over our collected dataset. Thus, generalization over other datasets is unknown, making comparisons across models difficult, and counts as a major limitation of this paper. Future work in this area would benefit greatly from corpora creation towards this end, and a diversity in languages and modalities would contribute greatly to the field.

Another area of future work comprises of experiments with $\text{HeterMPC}_{\text{BART}}$, also presented in Gu et al. (2022). The experiments showcased in this paper focus on $\text{HeterMPC}_{\text{BERT}}$ owing to timing and computation resource constraints. Thus, investigations into utilizing other architectures within the HeterMPC model, including its ability to utilize large pretrained language models (PLMs), form another limitation of this paper. Additionally, investigations into computational resources form another limitation for this paper, with GPUs required for model training for an acceptable time frame.

Lastly, there is a need for comparison with large language model (LLM) capabilities. However, multi-party support is not native for LLMs, and thus we focus on utilizing HeterMPC which allows us to model conversations natively. A future study focusing on adapting LLMs is planned future work, since it is outside the scope of this work.

Ethical Considerations

We recognize that there is potential for misuse based on our work. Persona-aware models have been shown in previous research (and ours) to perform more in keeping with the expected properties modeled by the system, which can make them seem more human and could be deceiving. Given the political nature of the dataset, there is potential to provoke emotional firestorms.

However, we also recognize that it could pave the way forward for more meaningful interactions in multi-party conversations on social media. Conversely, this research could provide a way for emotional regulation, and enhance discussions on social media by facilitating a more balanced conversation. It is our hope that this research is utilized in this direction.

Speaker		Addressee		Utterance		
ID	Persona	ID	Persona	ID	Parent ID	Text
1	white male independent expressor	-1	-	0	-	when you can't take a joke.... <link>
2	white male liberal expressor	1	white male independent expressor	1	0	this is a mix of toxic masculinity and privilege (rich / famous) on display. the joke was in poor taste - yes. but resorting to violence to d̈efendÿour wife from a joke. unacceptable. also, any other person (not rich / famous) would have been asked to leave / arrested.
3	white male liberal expressor	1	white male independent expressor	2	0	i can only hope that it was a staged event and not real.
4	white female conservative expressor	1	white male independent expressor	3	0	which is why ricky gervais will probably never host anything again.
5	white male liberal spectator	1	white male independent expressor	4	1	pathetic display by will smith (**Human**)
5	white male liberal spectator	1	white male independent expressor	4	1	i don't think we have to see how he're in this is in a lot of the country. (**HMPC**)
5	white male liberal spectator	1	white male independent expressor	4	1	. is a lot of them. (**PHMPC$_{\text{concat}}$**)
5	white male liberal spectator	1	white male independent expressor	4	1	it's not a good one! (**PHMPC$_{\text{graph}}$**)
5	white male liberal expressor	1	white male independent expressor	4	1	that's right.. (**PHMPC$_{\text{concat}}$**)
5	white male liberal expressor	1	white male independent expressor	4	1	it's not a lot of the same thing to be so, but they are so it. (**PHMPC$_{\text{graph}}$**)
5	white male liberal suppressor	1	white male independent expressor	4	1	that't....'s not just like a lot, i don's a lot of the same people who is. (**PHMPC$_{\text{concat}}$**)
5	white male liberal suppressor	1	white male independent expressor	4	1	it's just a lot of the real. (**PHMPC$_{\text{graph}}$**)
5	white male liberal avoider	1	white male independent expressor	4	1	is the time to do. (**PHMPC$_{\text{concat}}$**)
5	white male liberal avoider	1	white male independent expressor	4	1	they are right! (**PHMPC$_{\text{graph}}$**)

Table 7: Case study for comparing ground truth, and generated responses by HeterMPC (HMPC) & PersonaHeterMPC (PHMPC).

Acknowledgements

This research is part of a multi-phase study funded by the Department of Defense's Army Research Office through federal grant #72487-RT-REP.

7. Bibliographical References

Shubham Agarwal, Ondřej Dušek, Sebastian Gehrmann, Dimitra Gkatzia, Ioannis Konstas, Emiel Van Miltenburg, and Sashank Santhanam. 2020. Proceedings of the 1st workshop on evaluating nlg evaluation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*.

James E Allen, Curry I Guinn, and Eric Horvtz. 1999. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Maximillian Chen, Alexandros Papangelis, Chenyang Tao, Seokhwan Kim, Andy Rosenbaum, Yang Liu, Zhou Yu, and Dilek Hakkani-Tur. 2023. Places: Prompting language models for social conversation synthesis. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 844–868.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. Instructeval: Towards holistic evaluation of instruction-tuned large language models. *arXiv preprint arXiv:2306.04757*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Souvik Das, Sougata Saha, and Rohini K Srihari. 2023. Diving deep into modes of fact hallucinations in dialogue systems. *arXiv preprint arXiv:2301.04449*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Nouha Dziri, Sivan Milton, Mo Yu, Osmar R Zaiane, and Siva Reddy. 2022. On the origin of hallucinations in conversational models: Is it the datasets or the models? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5271–5285.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120. Association for Computational Linguistics.

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.

James J Gross. 1998. The emerging field of emotion regulation: An integrative review. *Review of general psychology*, 2(3):271–299.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware bert for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 2041–2044.

Jia-Chen Gu, Chao-Hong Tan, Chongyang Tao, Zhen-Hua Ling, Huang Hu, Xiubo Geng, and Daxin Jiang. 2022. Hetermpc: A heterogeneous graph neural network for response generation in multi-party conversations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5086–5097.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. Mpcbert: A pre-trained language model for multiparty 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 (Volume 1: Long Papers)*, pages 3682–3692.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

David Howcroft, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. 2020. Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definition. Association for Computational Linguistics (ACL).

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. Gsn: A graph-structured network for multi-party dialogues. In *International Joint Conference on Artificial Intelligence*.

Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *WWW'20: Proceedings of The Web Conference 2020*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Qi Jia, Yizhu Liu, Siyu Ren, Kenny Zhu, and Haifeng Tang. 2020. Multi-turn response selection using dialogue dependency relations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1911–1920.

Hang Jiang, Xianzhe Zhang, and Jinho D Choi. 2020. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract). In *Proceedings of the*

AAAI conference on artificial intelligence, volume 34, pages 13821–13822.

Dongshi Ju, Shi Feng, Pengcheng Lv, Daling Wang, and Yifei Zhang. 2022. Learning to improve persona consistency in multi-party dialogue generation via text knowledge enhancement. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 298–309.

Seokhwan Kim, Michel Galley, Chulaka Gunasekara, Sungjin Lee, Adam Atkinson, Baolin Peng, Hannes Schulz, Jianfeng Gao, Jinchao Li, Mahmoud Adada, et al. 2019. The eighth dialog system technology challenge. *arXiv preprint arXiv:1911.06394*.

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.

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 (Volume 1: Long Papers)*, 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 *International Joint Conference on Artificial Intelligence*.

Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. 2016. The teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431.

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.

Dayiheng Liu, Yu Yan, Yeyun Gong, Weizhen Qi, Hang Zhang, Jian Jiao, Weizhu Chen, Jie Fu, Linjun Shou, Ming Gong, et al. 2021. Glge: A new general language generation evaluation benchmark. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 408–420.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

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

Khyati Mahajan, Sourav Roy Choudhury, Sara Levens, Tiffany Gallicano, and Samira Shaikh. 2021. Community connect: A mock social media platform to study online behavior. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1073–1076.

Khyati Mahajan, Sashank Santhanam, and Samira Shaikh. 2022. Towards evaluation of multi-party dialogue systems. In *Proceedings of the 15th International Conference on Natural Language Generation*, pages 278–287.

Khyati Mahajan and Samira Shaikh. 2021. On the need for thoughtful data collection for multi-party dialogue: A survey of available corpora and collection methods. In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 338–352.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial intelligence review*, 56(4):3055–3155.

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*, pages 2133–2143.

Liang Qiu, Yizhou Zhao, Weiyan Shi, Yuan Liang, Feng Shi, Tao Yuan, Zhou Yu, and Song-chun Zhu. 2020. Structured attention for unsupervised dialogue structure induction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1889–1899.

Partha Pratim Ray. 2023. Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*.

Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tur. 2021. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. *arXiv preprint arXiv:2110.05456*.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Sakib Shahriar and Kadhim Hayawi. 2022. Let's have a chat! a conversation with chatgpt: Technology, applications, and limitations. In *Artificial Intelligence and Applications*.

Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11):992–1003.

Yizhou Sun, Brandon Norick, Jiawei Han, Xifeng Yan, Philip S Yu, and Xiao Yu. 2013. Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 7(3):1–23.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the twelfth ACM international conference on web search and data mining*, pages 267–275.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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.

Jimmy Wei, Kurt Shuster, Arthur Szlam, Jason Weston, Jack Urbanek, and Mojtaba Komeili. 2023. Multi-party chat: Conversational agents in group settings with humans and models. *arXiv preprint arXiv:2304.13835*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

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

Haisong Zhang, Zhangming Chan, Yan Song, Dongyan Zhao, and Rui Yan. 2018a. When less is more: Using less context information to generate better utterances in group conversations. In *Natural Language Processing and Chinese Computing: 7th CCF International Conference, NLPCC 2018, Hohhot, China, August 26–30, 2018, Proceedings, Part I 7*, pages 76–84. Springer.

Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. 2018b. Addressee and response selection in multi-party conversations with speaker interaction rnns. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

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

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