A Survey of Challenges and Methods in the Computational Modeling of Multi-Party Dialog

Ananya Ganesh Martha Palmer Katharina Kann
University of Colorado Boulder
ananya.ganesh@colorado.edu

Abstract

Advances in conversational AI systems, powered in particular by large language models, have facilitated rapid progress in understanding and generating dialog. Typically, task-oriented or open-domain dialog systems have been designed to work with two-party dialog, i.e., the exchange of utterances between a single user and a dialog system. However, modern dialog systems may be deployed in scenarios such as classrooms or meetings where conversational analysis of multiple speakers is required. This survey will present research around computational modeling of "multi-party dialog", outlining differences from two-party dialog, challenges and issues in working with multi-party dialog, and methods for representing multi-party dialog. We also provide an overview of dialog datasets created for the study of multi-party dialog, as well as tasks that are of interest in this domain.

1 Introduction

Dialog systems are increasingly a part of our personal and professional lives, and have made their way into domains such as healthcare (Valizadeh and Parde, 2022), business (Sang and Bao, 2022), and education (Litman and Silliman, 2004). Predominantly, research on dialog systems investigates how to develop task-oriented or open-domain systems that individual users can interact with, to accomplish routine tasks or engage in chit-chat. Conversations in such settings tend to be two-party or *dyadic* conversations, that is, involve only two participants, the system and the user, who may typically alternate turns while speaking. However, for applications such as classroom tutoring assistants or meeting summarization, dialog systems need to be able to understand and participate in *multi-party* dialog – interactions between multiple humans.

However, multi-party dialog is structurally different from dyadic dialog, requiring systems to be designed with their characteristics in mind. For

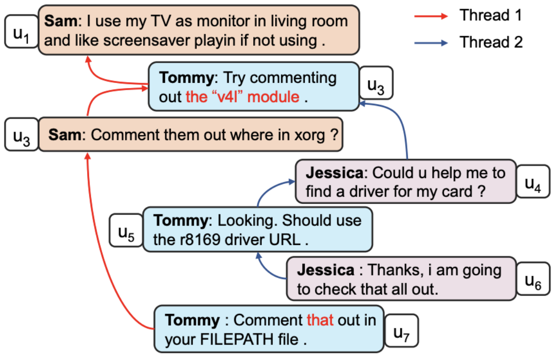

Figure 1: An example of a multi-party interaction, with speakers and threads marked. Figure from Shen et al. (2023)

instance, looking at the chat conversation in Figure 1, we see that the conversations are non-linear and interleaved, and utterances can be implicitly addressed to a specific participant(s). Conversational analysis of this interaction would require understanding each sub-dialog, and require resolving the speaker and addressees of each utterance. Responses by the dialog agent would also require determining which participant the response should be directed to. If multiple dialog agents are present, response management also requires determining which agent takes the turn. For the purposes of this study, we only consider scenarios with multiple human participants, and one dialog agent.

In this paper, we survey research that investigates the computational modeling of multi-party dialog [1]. We first introduce the characteristics of multi-party dialog based on early work in conversational analysis, focusing on ways in which they differ from two-party dialog. Based on these differences, we outline some of the challenges that face systems operating in this setting, and their solutions that have been investigated by the field. In Section 5, we present a comprehensive overview

[1]Unless stated otherwise, the systems and datasets we describe are focused on English dialog.

Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023), pages 140–154
July 14, 2023 ©2023 Association for Computational Linguistics

of representation learning methods for multi-party dialog, focusing on the merits of modeling information flow through graph structures, and discuss deep learning methods for obtaining and encoding these structures. Finally, we conclude with a discussion of opportunities for future work in multi-party dialog modeling.

2 Characteristics of Multi-Party Dialog

Participant Roles: The defining characteristic of multi-party dialog is the presence of multiple participants or interlocutors in a conversation. While in a two-party interaction, one participant takes on the role of the speaker in a turn and the other participant takes on the role of listener or "addressee", an utterance in a multi-party conversation not only has multiple candidate addressees, but could also be directed at multiple listeners at the same time. Traum (2004) further defines participant roles based on their degree of participation at various stages in the conversation: in-context listeners have heard all the previous utterances and may interpret the current utterance differently from a listener with no prior context; active participants are engaged in the conversation and play the roles of speakers and addressees, whereas overhearers may receive utterances but do not participate in the conversation.

Initiative and turn-taking: Traum (2004) observe that while many two-party dialog systems are mixed-initiative or user-initiative driven, multi-party dialog tends to be asymmetric in displaying initiative, with some participants dominating. Multi-party dialog may also include simultaneous conversations about multiple distinct topics (Elsner and Charniak, 2008). Aoki et al. (2006) analyze spontaneous social conversations in small groups, focusing on the nature of turn-taking in simultaneous conversations. Of particular interest are *conversational floors* (Sacks et al., 1974), which are structures that can be composed of one turn at a time such as in a therapy session, or can contain multiple alternating turns – for example, when a speaker has the floor and another speaker takes a turn to ask a question, but does not take the floor (Edelsky, 1981). They find that multi-party conversations tend to have multiple simultaneously active floors, with a single session (of up to an hour) having an average of 1.79 active floors, and a maximum of 4 active floors. They further find that floors are dynamic, particularly when the participants are young (ages 14-24) – in sessions with youth there are upto 70 distinct floors over the course of the conversation, each lasting about 44 seconds.

Dialog structure: Research has also studied how structures such as dialog acts or discourse relations can shed light on the nature of multi-party dialog. Ishizaki and Kato (1998) examine how dialog act structures differ between two-party and multi-party dialog (specifically, three-party dialog in their study). They first find that dialog act sequences most frequently involve only two speakers, particularly in sequences of length three to five. Looking at distances between utterances and their antecedents, Ginzburg and Fernández (2005) find that long range dependencies are more prevalent in multi-party dialog than in two-party dialog. Discourse relations prevalent in multi-party dialog also tend to be distinctive: Volha et al. (2011) find `feedback elicitation` to be more prevalent than in two-party dialog, whereas Asher et al. (2016) find that the most frequent relations are question-answer pairs or follow-up questions.

3 Challenges and Sub-Tasks

The unique characteristics of multi-party dialog imply the existence of challenges that cannot be handled by traditional two-party dialog systems. These challenges are occasionally treated as part of the larger system design (Ouchi and Tsuboi, 2016), but for the most part have been isolated as separate sub-tasks. We list a few major problems, and discuss solutions proposed in the literature.

3.1 Speaker and addressee recognition

In multi-party dialog, particularly in spoken or transcribed dialog, determining the speaker of the *current* utterance is a non-trivial task (Traum, 2004). *Closed-set* speaker identification is formulated as a classification task, where given an utterance, the goal is to determine the speaker from a list of known participants (Reynolds and Rose, 1995). Early work on text-independent speaker recognition makes use of acoustic features extracted from speech (Brunelli and Falavigna, 1995; Campbell et al., 2006) for classification, as well as multi-modal signals such as gestures (Bohus and Horvitz, 2010b) or the movement of lips in videos (Haider and Al Moubayed, 2012). Utterance-aware (Gu et al., 2022b) or text-dependent speaker identification uses the content of the utterance, typically from transcribed text, in order to determine the speaker. Work along these lines include Ma et al.

(2017), who classify speakers based on utterances from multiple transcripts and find success using a convolutional neural network, Meng et al. (2018) who use a hierarchical RNN (Serban et al., 2016) to encode content as well as temporal information indicated by speaker order.

Addressee identification is an important sub-task in which work follows two directions: 1) identifying the participant at whom each utterance is directed enables the construction of a graphical structure to represent information flow and 2) selecting the addressee to whom a response generated by a dialog agent should be addressed. For 1), Traum (2004) propose an algorithm looking at "vocative expressions" in the utterance, as well as speakers and content of current and previous utterances. Other features investigated for this task include gaze and acoustic features (Jovanovic et al., 2006; Jovanovic and op den Akker, 2004), and dialog acts (Gupta et al., 2007; Galley et al., 2004).

For 2), Ouchi and Tsuboi (2016) propose the task of *addressee and response selection*, where given a context of utterances with their speakers, the system predicts an addressee and a response. They propose two modeling frameworks, which both learn a vector representation for each participant (or agent), which is then encoded with the utterance context using an RNN: the *static* setting uses a fixed agent vector computed based on the speaking order of all agents, while the *dynamic* model updates the agent vector corresponding to the speaker of the current utterance at each timestep during training. However, since this doesn't capture the interaction between different agents, Zhang et al. (2018) propose an improvement that updates the embeddings of all active participants at each timestep. Wang et al. (2020) integrate addressee identification into a multi-task learning model that also performs topic prediction and response selection.

3.2 Turn taking

Turn-taking in natural conversations refers to the process by which humans coordinate participation, through verbal as well as non-verbal cues (Traum, 2004; Bohus and Horvitz, 2010b). Dialog systems, even in a two-party setting, need to perform turn management to identify when they can speak. Computational modeling of turn-taking in dialog is therefore a task that has received much attention (Hawes et al., 2009; Raux and Eskenazi, 2009; Bohus and Horvitz, 2010a; de Bayser et al., 2019). Bohus and Horvitz (2010a) define four kinds of "floor management" actions – *Hold, Release, Take* and *Null* to describe how turns move from one participant to another, and use heuristics based on response intervals to design a turn management system that chooses the appropriate action (Bohus and Horvitz, 2010b). Raux and Eskenazi (2009) use a similar formulation, and present a finite state machine that is optimized to minimize gaps and overlaps in a conversation.

Turn-taking is also modeled in some work as the task of predicting the next speaker, given a context consisting of speakers and utterances from previous turns. Hawes et al. (2009) treat this as a sequence labeling problem, and propose a second-order CRF in combination with features such as discourse markers (Marcu, 1997) and pronoun references. In more recent work, Skantze (2017) use lexical and acoustic features with an LSTM model; de Bayser et al. (2019) comparatively investigate SVM, CNN and LSTM models, achieving best results with the CNN models; Ishii et al. (2016) additionally use multi-modal features such as gaze to predict the next speaker as well as the time at which the next utterance will be made.

3.3 Conversation disentanglement

The presence of multiple simultaneous conversation floors (Section 2) results in distinct threads of conversation being entangled in a single session of multi-party dialogue. To enable understanding and responding to such conversations, the task of "conversation disentanglement" is important, which creates separate threads that are each about a specific topic. Elsner and Charniak (2008) introduce a corpus for this problem based on Internet Relay Chat (IRC) conversations, where annotations mark utterances that belong to the same conversational thread. They present a two-stage framework for disentanglement that first classifies pairs of utterances as to whether they are part of the same thread or not based on discourse and content features. Then, they perform correlation clustering to partition all utterances into clusters greedily. In follow-up work, Elsner and Charniak (2011) experiment with incorporating discourse coherence models (Lapata et al., 2005; Soricut and Marcu, 2006) for disentanglement, and find mixed results on the IRC corpus: models of local coherence help with assigning individual utterances into the right threads, but not in disentangling entire conversations.

The two-stage setup described here has been iteratively improved in future work, particularly by improving the classification component using deep learning models. Mehri and Carenini (2017) make use of discourse structure by annotating reply-to relations, and include two additional RNN-based classifiers to the Elsner and Charniak (2008) model, one for classifying pair-wise reply relations, and one for determining if an utterance follows a context. Jiang et al. (2018) achieve improvements to the same-thread classifier using Siamese CNNs. Kummerfeld et al. (2019) increase the scale of the IRC corpus by 30 times, creating a new benchmark for conversation disentanglement, and additionally propose an ensemble feedforward model that outperforms previous models. In contrast, more recent works investigate end-to-end models for this task, such as Liu et al. (2020) who develop a transition-based model that keeps track of states in discovered threads while assigning incoming utterances to existing or new threads in an online fashion. Liu et al. (2021) perform disentanglement on an unlabeled corpus by first creating pseudo data for the pairwise classifiers.

4 Datasets

Corpora for studying multi-party conversations span a variety of modalities – spoken (Renals et al., 2007), written (Lowe et al., 2015), or accompanied by video (Poria et al., 2019); they also span multiple genres, including chat forums for software discussions, movies and TV dialog, formal discourse in meetings and interviews, and informal discourse during gameplay. In this survey, we do not focus on comprehensively describing all available datasets, but provide an overview of three datasets which serve as benchmarks for modeling multi-party dialog, and have been extensively used in the models described below. For a detailed survey of datasets specifically, we refer the reader to Mahajan and Shaikh (2021).

Ubuntu IRC Corpora Internet Relay Chat (IRC), a text-based chat interface, contains channels for discussion about specialized topics. Typically, discussions consist of users posting questions, and other users replying with solutions, and all messages (or utterances), contain the identity of the sender (speaker). Corpora built from this interface have been used for the tasks of conversation disentanglement, speaker and addressee recognition, and response generation. Elsner and Charniak (2008) were the first to use conversations from the ##LINUX channel, which they manually annotate for threads, for the task of disentanglement. This yields 80 conversations, with a total of about 1500 utterances. Uthus and Aha (2013) scrape six years of chats from the ##ubuntu channel (which contains messages in English), as well as seven non-English channels including the languages Chinese, Russian, Spanish, Portuguese, Italian, Polish and Swedish. This corpus contains over 26 million messages, but without any annotations. Lowe et al. (2015) present the Ubuntu Dialog corpus, which contains 1 million English conversations totalling 7 million utterances. Each utterance contains speaker ID, and they also heuristically extract addressee IDs and disentangle conversations, as shown in Figure 2. Kummerfeld et al. (2019) present the largest manually annotated corpus from this domain, for the task of conversation disentanglement, with 70k utterances. Finally, Li et al. (2020) introduce the Molweni challenge corpus by annotating the Ubuntu corpus with reading comprehension style questions and answers, resulting in 33k question-answer pairs.

Time	User	Utterance
[12:21]	dell	well, can I move the drives?
[12:21]	cucho	dell: ah not like that
[12:21]	RC	dell: you can't move the drives
[12:21]	RC	dell: definitely not
[12:21]	dell	ok
[12:21]	dell	lol
[12:21]	RC	this is the problem with RAID:)
[12:21]	dell	RC haha yeah
[12:22]	dell	cucho, I guess I could just get an enclosure and copy via USB...
[12:22]	cucho	dell: i would advise you to get the disk

Sender	Recipient	Utterance
dell		well, can I move the drives?
cucho	dell	ah not like that
dell	cucho	I guess I could just get an enclosure and copy via USB
cucho	dell	i would advise you to get the disk
dell		well, can I move the drives?
RC	dell	you can't move the drives. definitely not. this is the problem with RAID :)
dell	RC	haha yeah

Figure 2: An interaction from Lowe et al. (2015), heuristically disentangled and tagged with addressees.

Meeting Corpora The AMI project (Kraaij et al., 2005; Renals et al., 2007) provides a corpus for multimodal conversational analysis of formal discourse – specifically, in multi-party meetings. The AMI corpus consists of 100 hours (175 sessions) of scenario-oriented meetings between four participants, where video and audio are recorded, along with artifacts such as digital pen movements and whiteboard content. They providing access to videos, manually transcribed speech, abstractive and extractive summaries of the conversations, and annotations for dialog acts, topic segments, gaze and positional information, and gestures. Other corpora under the umbrella of the AMI project includes the ICSI corpus (Janin et al., 2003), which contains 72 hours of naturally-occuring meetings (not elicited by a scenario).

MELD Corpus Another multi-modal multi-party dataset that is widely used in the models below is the MELD corpus (Poria et al., 2019), designed for emotion recognition from conversations. It consists of 1433 conversations from the TV show Friends, providing access to video, audio, and transcripts. They include annotations at the utterance level indicating one out of seven emotions (such as anger, surprise, etc.) expressed by the utterance.

5 Representation Learning for MPD

In this section, we will describe how machine learning models represent and encode multi-party dialog in order to leverage its inherent structural properties for tasks such as response generation. Early work such as Lowe et al. (2015) represent the entire conversational context sequentially, where all prior utterances to the current one that fall in a window are concatenated. Improvements such as Zhou et al. (2016) model relationships between the current utterance and the context through a hierarchical RNN. However, given that multi-party dialog can have multiple addressees, multiple replies, as well as simultaneous conversations, such sequential structures cannot represent all relationships between utterances in the dialog.

As a solution, recent successful models experiment with graph structures to represent the flow of information in multi-party dialog. Typically, this approach treats the utterances as nodes, and the relations between them (such as *reply-to*) as edges. The graphs thus obtained are encoded through a suitable neural network architecture (Kipf and Welling, 2017; Schlichtkrull et al., 2018), and the resulting embeddings are used for the downstream task, in combination with decoders or classification layers. Below, we look at specific sub-components and strategies for this workflow.

5.1 Dialog structure induction

Corpora such as the Ubuntu Dialog Corpus (Lowe et al., 2015), which serve as benchmarks for modeling multi-party dialog, contain explicit annotations for speakers and addressees. When annotations for dialog structure such as addressee information are not available, dialog structure needs to be learned from the conversation without explicit supervision, so that it can be used to perform downstream tasks While unsupervised methods for structure induction on task-oriented dialog have received some attention (Shi et al., 2019; Sun et al., 2021a; Xu et al., 2021), comparatively less work exists for multi-party dialog, the most prominent being Qiu et al. (2020), who propose a model to induce structure on both two-party and multi-party dialog. They propose a model for response generation, which consists of a Variational Recurrent Neural Network (VRNN) (Chung et al., 2015) into which *structured attention* layers are integrated, such that the latent state of the VRNN captures the underlying dialog structure. The model first encodes sentences with an LSTM, then the VRNN encodes a dialog history into a latent state, which is then decoded to produce a response. While training, they maximize the conditional likelihood of a response given the history, while also learning a latent dependency tree – here, nodes represents the utterances, and directed edges exist between nodes when one utterance is the parent of another. Evaluating on the Ubuntu Chat Corpus (Uthus and Aha, 2013), they find that the VRNN model performs comparably to a graph-based model that makes use of explicit speaker/addressee annotations (Hu et al., 2019). On comparing the learned utterance dependency tree with gold annotations for speaker and addressee relations, they find that the model achieves an accuracy of 68.5% in identifying the parents of each utterance.

5.2 Graph-based representations

Unlike Qiu et al. (2020), the predominant line of research on modeling multi-party dialog makes use of annotated speaker/addressee information in order to obtain the graph structures. Hu et al. (2019) propose a model for response generation that they call *Graph Structured Networks* (GSN), which was to our knowledge the first to successfully apply graphs to multi-party dialog. Similar to the framework discussed above, they formulate their graph as an utterance dependency graph, assuming access to annotated speaker/addressee information within the conversational data. The GSN consists of a word-level encoder to represent utterances, an utterance-level graph structured encoder to represent information flow, and a decoder to generate responses. Embeddings for an utterance are obtained from the graph using forward and backward information flow, and the speaker information. In experiments on the Ubuntu Dialog Corpus (Lowe et al., 2015), they find that their proposed model achieves a significant improvement over baselines that are based on sequential or hierarchical utterance encodings (Serban et al., 2016). They further find, through ablations, that the inclusion of speaker information flow is crucial to model performance.

For two-party and task-oriented dialog, Graph Convolutional Networks (Kipf and Welling, 2017; Schlichtkrull et al., 2018) have been successfully used for representing structure (Banerjee and Khapra, 2019), and have consequently been explored for multi-party dialog as well. Ghosal et al. (2019) propose a model called DialogueGCN for the task of emotion recognition from conversations, which is an utterance-level classification task. They represent each utterance as a node in the graph, and construct edges to represent the context – all utterances within a window prior and after the current utterance are marked. They also assign relational edges, to capture temporal dependency as well as speaker dependency between pairs of utterances. The graph is then encoded through Relational Graph Convolutional Networks (Schlichtkrull et al., 2018), which provides a representation for each node that aggregates information from its context nodes. The proposed model outperforms multiple strong baselines when evaluating on MELD (Poria et al., 2019), including DialogRNNs (Majumder et al., 2019). A similar framework is proposed by Ju et al. (2022), who include *personas* corresponding to each speaker in the vertex set, for the task of generating personalized responses. Edges are then constructed between personas and their corresponding utterances, as well as between consecutive utterances, before encoding through a GCN. As a baseline, they adapt DialogueGCNs for response generation by adding a decoder, and show the superiority of their persona-aware model according to automated and human evaluation metrics.

Similar to Ju et al. (2022), the idea of including nodes that are not just utterances has been explored by other work, resulting in graphs that are *heterogenous*. Gu et al. (2022a) propose *HeterMPC*, a graph-based model for response generation in multi-party dialog. Their graph treats utterances as well as participants as nodes, drawing edges between nodes to indicate six types of relations: *reply, reply-to, speak, spoken-by, address, addressed-by*. Utterance nodes are represented by embeddings from BERT, whereas interlocutors are represented by a speaker embedding initialized based on their position in the conversation. When updating the representations for nodes, they compute heterogeneous attention weights over source and target, conditioned on the edge type. Their proposed model outperforms GSNs with automated and human evaluations. Further, their ablations indicate the importance of interlocutor nodes as well as edge relations. Sang and Bao (2022) also make use of heterogeneous graphs that contain participant and utterance nodes, towards the task of financial risk prediction upon earnings call conferences. The edges in their graph connect speakers to their utterances, and the resulting graph is encoded with a Graph Attention Network (Veličković et al., 2018). From the graph encoder's output, they aggregate speaker embeddings separately from utterance embeddings using two separate contextual attention layers, which then represent the whole conversation, which is then classified for stock volatility. Lee and Choi (2021) include four types of nodes in their graph: dialog (utterance), turn, subject, and object; edges relate turns nodes to their respective utterances, connect utterances by the same speaker, and connect turns to arguments that are mentioned. They also encode their graph with a GCN, and evaluate on the tasks of relation extraction in dialogues, as well as emotion recognition. Liang et al. (2021) take heterogeneous graphs one step further with multimodal nodes – their nodes include utterances, facial expression features, emotion categories, and speakers, with seven kinds of edges capturing the relations between the different features. They encode this graph with a heterogeneous graph neural network (Zhang et al., 2019), and evaluate on the downstream task of response generation expressing a suitable emotion.

5.3 Utilizing discourse relations

Some research has investigated how the graph structures described above can include other task-specific or linguistic information, such as annotations for discourse.

Feng et al. (2021) present a dialog discourse aware graph-based model for the task of meeting summarization. Of interest are 16 discourse relations from Asher et al. (2016) including comment, QA, elaboration, etc. They obtain discourse relations from a dialog discourse parser (Shi and Huang, 2019), and transform it such that nodes are created for utterances as well as discourse relations, with directed edges marking the relations between utterances. They encode their graph with an R-GCN (Schlichtkrull et al., 2018). Experiments on the AMI and IMSI meeting corpora show improvements over sequential models (Serban et al., 2016). They find that performance is correlated with the quality of the discourse parser, as well as the number of discourse relations available. Discourse structures from an off-the-shelf parser are also used by Sun et al. (2021b) in their graph-based model for emotion recognition. Similar to Ghosal et al. (2019), they construct directed edges between utterance nodes, marking discourse relations in addition to speaker and temporal relations. The inclusion of discourse results in a significant improvement over DialogGCNs on the MELD corpus. Contemporaneously, Li et al. (2021) investigate discourse-aware graphs for machine reading comprehension on multi-party dialog as found in the Molweni challenge corpus (Li et al., 2020). They also model utterances as nodes, with dependencies as edges and discourse types denoted by edge relations, using DialogGCN for encoding. Additionally, an MRC module integrates a representation for the question, outputting an answer span.

5.4 Pretraining

Following the advancements in the representational capabilities of pretrained language models (Devlin et al., 2019; Radford and Narasimhan, 2018), models such as ToD-BERT (Wu et al., 2020) and DialoGPT (Zhang et al., 2020) have been developed with the goal of enhancing dialog representations in task-oriented or open-domain dialog. Pre-training has also been explored for multi-party dialog: Gu et al. (2021) propose MPC-BERT, in which they pre-train BERT on data from the Ubuntu Chat Corpus (Lowe et al., 2015), with five self-supervision tasks. These tasks are designed to model underlying interlocutor structure in multi-party dialog, as well as utterance semantics. Tasks for the first category include 1) *reply-to utterance recognition*, which involves predicting the preceding utterance that an utterance is replying to; 2) *identical speaker searching*, or identifying utterances that share a speaker; 3) *pointer-consistency distinction*, which involves maintaining a similar representation for pairs of utterances between the same speaker–addressee pair in order to model interlocutors. Tasks for the second category include 1) *masked shared utterance restoration*, where utterances that receive multiple replies are masked and reconstructed during training 2) *shared node detection*, where sub-threads of the same parent utterance are required to be correctly identified. The pretrained model thus obtained can be finetuned for downstream tasks – the authors finetune and evaluate on the tasks of addressee recognition, speaker identification, and response selection, outperforming previous methods significantly. Notably, all of the finetuning tasks are from the same domain (Ubuntu IRC) as the pre-training data, although the authors declare that they only use the train split for pre-training.

Other work that focuses on pre-training for multi-party conversation understanding includes Zhong et al. (2022), who focus on learning long-range dependencies across dialog, in order to solve problems like summarization and question answering. In contrast to MPC-BERT, and similar to BART (Lewis et al., 2019), their self-supervision objective involves denoising dialog based on windows – given a long dialog, they sample random windows to which noise is added, which is later reconstructed. The added noise takes the form of masking speaker identities, utterances, merging turns and shuffling utterances within a turn. With this objective, they train a Transformer-based model called UniLM (Dong et al., 2019) on the Movie Subtitles corpus (Lison and Tiedemann, 2016) and MediaSum interview corpus (Zhu et al., 2021). Finetuning on the tasks of summarization, dialog segmentation and question answering, they show improvements across automated and human evaluations. Wang et al. (2020) pretrain a BERT model on the task of topic prediction – determining if two utterances are about the same topic, in addition to masked language modeling. Their encoder, called TopicBERT, is then finetuned in a multi-task learning setup, on the tasks of response selection, topic prediction, and topic disentanglement.

6 Tasks of Interest

Response generation and selection: As seen above, a large body of work exists on response generation (Qiu et al., 2020; Hu et al., 2019; Gu et al., 2022a), given a multi-party dialog as context. To generate responses at the right time and towards the right speaker, this can be combined with the tasks of speaker prediction (Yang et al., 2019) and addressee selection (Liu et al., 2019). The generated responses are typically evaluated with a combination of automated metrics such as BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) given a reference from the conversation. Human evaluations, such as in Liu et al. (2019); Gu et al. (2022a); Ju et al. (2022) assess whether responses are fluent, consistent with the context, informative, and coherent. The task of response selection, formulated as *retrieving* the most appropriate next utterance from a set of candidates, is also of interest (Ouchi and Tsuboi, 2016; Zhang et al., 2018; Wang et al., 2020; Gu et al., 2021). Response selection is typically evaluated with classification-based metrics such as precision and recall, including $Recall_n@k$ to match n available candidates with top k retrieved candidates.

Modeling socio-cultural phenomena: Multi-party conversations are of interest from a computational social science perspective, to study interactional dynamics between participants. This includes determining when decision-making occurs (Frampton et al., 2009; Bui et al., 2009), analyzing bargaining and negotiation strategies (Petukhova et al., 2016; Joshi et al., 2021; Asher et al., 2016), and analyzing collaborative behavior such as entrainment (Litman et al., 2016; Rahimi et al., 2017), cohesion (Bangalore Kantharaju et al., 2020) and agreement (Hillard et al., 2003; Strzalkowski et al., 2010; Rosenthal and McKeown, 2015). Work on recognizing emotions from utterances, typically with multi-modal information, is also loosely related to this direction (Ghosal et al., 2019; Poria et al., 2019).

Other NLP tasks: Datasets and models have been developed for the task of summarization of multi-party conversations (Renals et al., 2007; Purver et al., 2007; Chen and Metze, 2012; Zhu et al., 2021). While Zhu et al. (2021) provide a dataset that disentangles the primary topic from secondary topics before summarization, an underexplored issue is performing summarization jointly with disentanglement so that multiple summaries are produced for the multiple sub-threads in the conversation. Other high-level NLP tasks that have been explored include answering reading comprehension questions over multi-party dialog (Li et al., 2020, 2021), and relation extraction (Albalak et al., 2022; Yu et al., 2020).

7 Discussion

One of the salient findings from our survey is that most recent work on multi-party dialog modeling, particularly using the graph-based methods, are centered around corpora from a limited set of domains; in fact, almost all of the models in Section 5 are evaluated on the Ubuntu chat corpus or on TV show transcript corpora. A possible reason for this is the availability of annotated structure in these datasets, including speaker and addressee information, as well as threads. However, we argue that the time is ripe for researchers to investigate how to extend modeling innovations to other available corpora and domains.

This is an important next step for two reasons, namely *real-world applicability*, and *robustness*. Natural dialog, such as spontaneous interactions between humans, is typically not well-represented in datasets such as typed chat, or scripted TV dialog. With the growing influence of dialog systems in daily lives, if our goal is to build better technology for the real world, like classrooms or businesses, we need to demonstrate that these state-of-the-art models perform equally well on probable, real-world conversations. Moreover, as seen in Mahajan and Shaikh (2021), numerous datasets satisfying these properties are actually available, although they do not necessarily contain explicit annotations for structure. However, as this survey shows, we have a large body of work that tells us how to go from natural conversations to more structured representations through tasks such as speaker and addressee recognition, turn prediction, and conversation disentanglement. Using these tasks as scaffolds for downstream tasks like response generation would enable us to leverage the expressivity of graph-based modeling on new and realistic domains.

In terms of other important next steps for this field of research, one interesting direction is exploring strategies for obtaining silver-standard graph structures through unsupervised methods – we so far only find one paper constructing a reply-to relation graph unsupervisedly. Additionally, to answer the robustness question, a systematic assessment of the advantages and shortcomings of graph-structured methods on rarer domains such as meetings (Petukhova et al., 2016) could be highly valuable, particularly for practitioners interested in studying the phenomena exhibited in such conversations. More broadly in this direction, given how the methods we have seen are predominantly focused on English multi-party dialog, the applicability of these methods to languages other than English (Liu et al., 2012), as well as conversations with code-switching (Hartmann et al., 2018), also needs to be evaluated. Finally, with the growing adoption and effectiveness of large language models (LLMs) in NLP research, a natural next question is to determine how these models can be used in understanding multi-party dialog, and what their limitations are. Current directions with promising results include using LLMs for conversation synthesis (Wei et al., 2023; Chen et al., 2023), where high-quality multi-party conversations are synthesized through prompting, and the conversations can be grounded in specific characters or personas. Such synthesized conversations may also help adapt methods for conversation analysis and response generation to rarer domains that may not be well-represented in natural corpora.

8 Conclusion

Our survey provides an overview of research in computationally modeling multi-party dialog. We identify major challenges based on differences from two-party dialog, and discuss how sub-tasks have been designed for solving them. We comprehensively describe recent advances in representation learning for multi-party dialog, focusing in particular on graph-based structures. Finally, we discuss some key directions that future work in this area can explore.

Acknowledgments

We thank the anonymous reviewers for their thoughtful feedback and suggestions. We also thank the members of the CU Boulder Computational Semantics group for their feedback on this survey. This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of the NSF.

References

Alon Albalak, Varun Embar, Yi-Lin Tuan, Lise Getoor, and William Yang Wang. 2022. D-REX: Dialogue relation extraction with explanations. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 34–46, Dublin, Ireland. Association for Computational Linguistics.

Paul M. Aoki, Margaret H. Szymanski, Luke Plurkowski, James D. Thornton, Allison Woodruff, and Weilie Yi. 2006. Where's the "party" in "multi-party"? analyzing the structure of small-group sociable talk. In *Proceedings of the 2006 20th Anniversary Conference on Computer Supported Cooperative Work*, CSCW '06, page 393–402, New York, NY, USA. Association for Computing Machinery.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Suman Banerjee and Mitesh M Khapra. 2019. Graph convolutional network with sequential attention for goal-oriented dialogue systems. *Transactions of the Association for Computational Linguistics*, 7:485–500.

Reshmashree Bangalore Kantharaju, Caroline Langlet, Mukesh Barange, Chloé Clavel, and Catherine Pelachaud. 2020. Multimodal analysis of cohesion in multi-party interactions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 498–507, Marseille, France. European Language Resources Association.

Dan Bohus and Eric Horvitz. 2010a. Computational models for multiparty turn taking. *Technical Report. Microsoft Research Technical Report MSR-TR 2010-115*.

Dan Bohus and Eric Horvitz. 2010b. Facilitating multiparty dialog with gaze, gesture, and speech. In *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, pages 1–8.

Roberto Brunelli and Daniele Falavigna. 1995. Person identification using multiple cues. *IEEE transactions on pattern analysis and machine intelligence*, 17(10):955–966.

Trung Bui, Matthew Frampton, John Dowding, and Stanley Peters. 2009. Extracting decisions from multi-party dialogue using directed graphical models and semantic similarity. In *Proceedings of the SIGDIAL 2009 Conference*, pages 235–243, London, UK. Association for Computational Linguistics.

William M Campbell, Joseph P Campbell, Douglas A Reynolds, Elliot Singer, and Pedro A Torres-Carrasquillo. 2006. Support vector machines for speaker and language recognition. *Computer Speech & Language*, 20(2-3):210–229.

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, Dubrovnik, Croatia. Association for Computational Linguistics.

Yun-Nung Chen and Florian Metze. 2012. Intra-speaker topic modeling for improved multi-party meeting summarization with integrated random walk. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 377–381, Montréal, Canada. Association for Computational Linguistics.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. 2015. A recurrent latent variable model for sequential data. *Advances in neural information processing systems*, 28.

Maíra Gatti de Bayser, Paulo Rodrigo Cavalin, Claudio Santos Pinhanez, and Bianca Zadrozny. 2019. Learning multi-party turn-taking models from dialogue logs. *CoRR*, abs/1907.02090.

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

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. *Advances in neural information processing systems*, 32.

Carole Edelsky. 1981. Who's got the floor? *Language in Society*, 10(3):383–421.

Micha Elsner and Eugene Charniak. 2008. You talking to me? a corpus and algorithm for conversation disentanglement. In *Proceedings of ACL-08: HLT*, pages 834–842, Columbus, Ohio. Association for Computational Linguistics.

Micha Elsner and Eugene Charniak. 2011. Disentangling chat with local coherence models. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1179–1189.

Xiachong Feng, Xiaocheng Feng, Bing Qin, and Xinwei Geng. 2021. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3808–3814. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Matthew Frampton, Jia Huang, Trung Bui, and Stanley Peters. 2009. Real-time decision detection in multi-party dialogue. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1133–1141, Singapore. Association for Computational Linguistics.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. 2004. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 669–676, Barcelona, Spain.

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, Hong Kong, China. Association for Computational Linguistics.

Jonathan Ginzburg and Raquel Fernández. 2005. Scaling up from dialogue to multilogue: Some principles and benchmarks. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 231–238, Ann Arbor, Michigan. Association for Computational Linguistics.

Jia-Chen Gu, Chao-Hong Tan, Chongyang Tao, Zhen-Hua Ling, Huang Hu, Xiubo Geng, and Daxin Jiang. 2022a. 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, Dublin, Ireland. Association for Computational Linguistics.

Jia-Chen Gu, Chongyang Tao, and Zhen-Hua Ling. 2022b. Who says what to whom: A survey of multi-party conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 5486–5493.

Jia-Chen Gu, Chongyang Tao, Zhenhua Ling, Can Xu, Xiubo Geng, and Daxin Jiang. 2021. 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 (Volume 1: Long Papers)*, pages 3682–3692, Online. Association for Computational Linguistics.

Surabhi Gupta, John Niekrasz, Matthew Purver, and Dan Jurafsky. 2007. Resolving “you” in multi-party dialog. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 227–230, Antwerp, Belgium. Association for Computational Linguistics.

Fasih Haider and Samer Al Moubayed. 2012. Towards speaker detection using lips movements for human-machine multiparty dialogue. In *The XXVth Swedish Phonetics Conference (FONETIK)*, pages 117–120. Citeseer.

Silvana Hartmann, Monojit Choudhury, and Kalika Bali. 2018. An integrated representation of linguistic and social functions of code-switching. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Timothy Hawes, Jimmy Lin, and Philip Resnik. 2009. Elements of a computational model for multi-party discourse: The turn-taking behavior of supreme court justices. *Journal of the American Society for Information Science and Technology*, 60(8):1607–1615.

Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. 2003. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Companion Volume of the Proceedings of HLT-NAACL 2003 - Short Papers*, pages 34–36.

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 Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5010–5016. International Joint Conferences on Artificial Intelligence Organization.

Ryo Ishii, Kazuhiro Otsuka, Shiro Kumano, and Junji Yamato. 2016. Prediction of who will be the next speaker and when using gaze behavior in multiparty meetings. *ACM Trans. Interact. Intell. Syst.*, 6(1).

Masato Ishizaki and Tsuneaki Kato. 1998. Exploring the characteristics of multi-party dialogues. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 583–589, Montreal, Quebec, Canada. Association for Computational Linguistics.

Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al. 2003. The icsi meeting corpus. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03).*, volume 1, pages I–I. IEEE.

Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. 2018. Learning to disentangle interleaved conversational threads with a Siamese hierarchical network and similarity ranking. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1812–1822, New Orleans, Louisiana. Association for Computational Linguistics.

Rishabh Joshi, Vidhisha Balachandran, Shikhar Vashishth, Alan W. Black, and Yulia Tsvetkov. 2021. Dialograph: Incorporating interpretable strategy-graph networks into negotiation dialogues. *CoRR*, abs/2106.00920.

Natasa Jovanovic and Rieks op den Akker. 2004. Towards automatic addressee identification in multi-party dialogues. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 89–92, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

Natasa Jovanovic, Rieks op den Akker, and Anton Nijholt. 2006. Addressee identification in face-to-face meetings. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 169–176, Trento, Italy. Association for Computational Linguistics.

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, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus.

Jonathan K. Kummerfeld, Sai R. Gouravajhala, Joseph J. Peper, Vignesh Athreya, Chulaka Gunasekara, Jatin Ganhotra, Siva Sankalp Patel, Lazaros C Polymenakos, and Walter Lasecki. 2019. A large-scale corpus for conversation disentanglement. In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics, pages 3846–3856, Florence, Italy. Association for Computational Linguistics.

Mirella Lapata, Regina Barzilay, et al. 2005. Automatic evaluation of text coherence: Models and representations. In *Ijcai*, volume 5, pages 1085–1090.

Bongseok Lee and Yong Suk Choi. 2021. Graph based network with contextualized representations of turns in dialogue. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 443–455, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

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

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

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*, volume 35, pages 13343–13352.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

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, Austin, Texas. Association for Computational Linguistics.

Diane Litman and Scott Silliman. 2004. Itspoke: An intelligent tutoring spoken dialogue system. In *Demonstration papers at HLT-NAACL 2004*, pages 5–8.

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, Hong Kong, China. Association for Computational Linguistics.

Hui Liu, Zhan Shi, Jia-Chen Gu, Quan Liu, Si Wei, and Xiaodan Zhu. 2020. End-to-end transition-based online dialogue disentanglement. In *IJCAI*, volume 20, pages 3868–3874.

Hui Liu, Zhan Shi, and Xiaodan Zhu. 2021. Unsupervised conversation disentanglement through co-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2345–2356, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ting Liu, Samira Shaikh, Tomek Strzalkowski, George Aaron Broadwell, Jennifer Stromer-Galley, Sarah M Taylor, Umit Boz, Xiaoai Ren, and Jingsi Wu. 2012. Extending the mpc corpus to chinese and urdu-a multiparty multi-lingual chat corpus for modeling social phenomena in language. In *LREC*, pages 2868–2873.

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

Kaixin Ma, Catherine Xiao, and Jinho D. Choi. 2017. Text-based speaker identification on multiparty dialogues using multi-document convolutional neural networks. In *Proceedings of ACL 2017, Student Research Workshop*, pages 49–55, Vancouver, Canada. Association for Computational Linguistics.

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, Singapore and Online. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6818–6825.

Daniel Marcu. 1997. From discourse structures to text summaries. In *Intelligent Scalable Text Summarization*.

Shikib Mehri and Giuseppe Carenini. 2017. Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 615–623, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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 Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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, Austin, Texas. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Volha Petukhova, Christopher Stevens, Harmen de Weerd, Niels Taatgen, Fokie Cnossen, and Andrei Malchanau. 2016. Modelling multi-issue bargaining dialogues: Data collection, annotation design and corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3133–3140, Portorož, Slovenia. European Language Resources Association (ELRA).

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. 2007. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 18–25, Antwerp, Belgium. Association for Computational Linguistics.

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, Online. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Zahra Rahimi, Anish Kumar, Diane Litman, Susannah Paletz, and Mingzhi Yu. 2017. Entrainment in Multi-Party Spoken Dialogues at Multiple Linguistic Levels. In *Proc. Interspeech 2017*, pages 1696–1700.

Antoine Raux and Maxine Eskenazi. 2009. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of human language technologies: The 2009 annual conference of the North American chapter of the association for computational linguistics*, pages 629–637.

Steve Renals, Thomas Hain, and Hervé Bourlard. 2007. Recognition and understanding of meetings the ami and amida projects. In *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 238–247. IEEE.

Douglas A Reynolds and Richard C Rose. 1995. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE transactions on speech and audio processing*, 3(1):72–83.

Sara Rosenthal and Kathy McKeown. 2015. I couldn't agree more: The role of conversational structure in agreement and disagreement detection in online discussions. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 168–177, Prague, Czech Republic. Association for Computational Linguistics.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Yunxin Sang and Yang Bao. 2022. DialogueGAT: A graph attention network for financial risk prediction by modeling the dialogues in earnings conference calls. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1623–1633, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web: 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings 15*, pages 593–607. Springer.

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.

Weizhou Shen, Xiaojun Quan, and Ke Yang. 2023. Generic dependency modeling for multi-party conversation. *ArXiv*, abs/2302.10680.

Weiyan Shi, Tiancheng Zhao, and Zhou Yu. 2019. Unsupervised dialog structure learning. 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 1797–1807, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7007–7014.

Gabriel Skantze. 2017. Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 220–230, Saarbrücken, Germany. Association for Computational Linguistics.

Radu Soricut and Daniel Marcu. 2006. Discourse generation using utility-trained coherence models. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 803–810.

Tomek Strzalkowski, George Aaron Broadwell, Jennifer Stromer-Galley, Samira Shaikh, Sarah Taylor, and Nick Webb. 2010. Modeling socio-cultural phenomena in discourse. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1038–1046, Beijing, China. Coling 2010 Organizing Committee.

Yajing Sun, Yong Shan, Chengguang Tang, Yue Hu, Yinpei Dai, Jing Yu, Jian Sun, Fei Huang, and Luo Si. 2021a. Unsupervised learning of deterministic dialogue structure with edge-enhanced graph autoencoder. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 13869–13877.

Yang Sun, Nan Yu, and Guohong Fu. 2021b. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2949–2958, Punta Cana, Dominican Republic. Association for Computational Linguistics.

David Traum. 2004. Issues in multiparty dialogues. In *Advances in Agent Communication: International Workshop on Agent Communication Languages, ACL 2003, Melbourne, Australia, July 14, 2003. Revised and Invited Papers*, pages 201–211. Springer.

David C Uthus and David W Aha. 2013. The ubuntu chat corpus for multiparticipant chat analysis. Technical report, NAVAL RESEARCH LAB WASHINGTON DC.

Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660, Dublin, Ireland. Association for Computational Linguistics.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Petukhova Volha, Laurent Prevot, and Bunt Harry. 2011. Multi-level discourse relations between dialogue units. In *The Sixth Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation*, pages 18–27.

Weishi Wang, Steven C.H. 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, Online. Association for Computational Linguistics.

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.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Jun Xu, Zeyang Lei, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. Discovering dialog structure graph for coherent dialog generation. 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 1726–1739.

Qichuan Yang, Zhiqiang He, Zhiqiang Zhan, Jianyu Zhao, Yang Zhang, and Changjian Hu. 2019. Mids: End-to-end personalized response generation in untrimmed multi-role dialogue. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Dian Yu, Kai Sun, Claire Cardie, and Dong Yu. 2020. Dialogue-based relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4927–4940, Online. Association for Computational Linguistics.

Chuxu Zhang, Dongjin Song, Chao Huang, Ananthram Swami, and Nitesh V Chawla. 2019. Heterogeneous graph neural network. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 793–803.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Ming Zhong, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2022. Dialoglm: Pre-trained model for long dialogue understanding and summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11765–11773.

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381, Austin, Texas. Association for Computational Linguistics.

Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. 2021. MediaSum: A large-scale media interview dataset for dialogue summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5927–5934, Online. Association for Computational Linguistics.