TOWARDS MULTI-PARTY CONVERSATION MODELING

by

Khyati Mahajan

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Computing and Information Systems

Charlotte

2023

Approved by:

Dr. Samira Shaikh

Dr. Wlodek Zadrozny

Dr. Razvan Bunescu

Dr. Xi (Sunshine) Niu

Dr. Mohamed Shehab

ABSTRACT

KHYATI MAHAJAN. Towards Multi-Party Conversation Modeling. (Under the direction of DR. SAMIRA SHAIKH)

Recent advances in the field of Natural Language Processing, specifically in Natural Language Generation (NLG) towards Dialogue Systems have focused mainly on two-party conversations. However, group conversations or multi-party conversations (MPC) are just as prevalent in our everyday lives. While the area of multi-party conversation modeling has received some attention in recent times, MPC lacks resources for 1) corpora in differing settings (formal/informal, synchronous/asynchronous), 2) dialogue models which can participate in informal open-domain settings while maintaining speaker information, and 3) evaluation metrics which provide better insights into the performance of MPC models when it comes to operating in groups and interacting with multiple participants. We thus take a three-pronged approach towards contributing to research in the MPC modeling research area. For corpora collection, we contribute a mock social media tool that can be utilized for collecting asynchronous MPC conversations called Community Connect, and utilize it for three separate experiments to collect everyday talk. Utilizing this tools also allows us to obtain informed consent. For MPC modeling, we propose a response generation model, using large language models (LLMs) and graph structured networks, which is capable of taking participant relations into account towards maintaining multiple persona profiles and generating responses keeping the speaker characteristics in mind. We find that this persona-aware response generation performs better than the baseline model. Lastly, for MPC evaluation, we present an expansion to the taxonomy of errors which contributes MPC-specific metrics to the overall NLG errors. In addition to the taxonomy, we contribute to better evaluation standards across which progress in the tasks within MPC can be tracked more saliently. Through these contribu-

tions, we aim to fill the necessary gaps towards advancing MPC understanding and modeling, while also providing the tools to gauge progress until now.

DEDICATION

I dedicate this dissertation to my great support system, who have stood by me and provided unwavering support, guidance, and encouragement throughout this long academic journey. Your belief in me has been instrumental in the successful completion of this work. Much of my PhD journey (and consequently this work) was conducted during the COVID-19 pandemic, which led to innumerable unforeseen obstacles which you all have supported me through. The delays in completing research and user studies, changes in grants and the way we communicated with the world, and the inability to attend conferences in person were all major obstacles to mention the least, and I could not have overcome them alone without your help.

To my esteemed advisor, Dr. Samira Shaikh, your guidance and mentorship have been invaluable. Your expertise, patience, and dedication to pushing the boundaries of knowledge have inspired me to reach new heights. Your insightful feedback, constructive criticism, and endless encouragement have shaped me into the researcher I am today. I am grateful for your unwavering support and the countless hours you have invested in helping me bring this dissertation to fruition.

To my loving parents, Shweta Mahajan and Chandrashekhar Mahajan, thank you for your unwavering love, and belief in my abilities. You have been my pillars of strength, providing constant support and encouragement throughout my life. I am forever grateful for the sacrifices you have made to ensure my success, and I dedicate this work to you both. To my late aba Sripad Shivapurkar and my lovely aaji Pushpa Shivapurkar, I am forever indebted to your support and faith in my abilities. Not a day goes by that I wish you could read this dissertation, aba.

To my partner Parth Shah, who has stood by my side through thick and thin; thank you for your support, understanding, and encouragement. Your belief in my abilities and your willingness to lend an ear or a helping hand whenever needed have been immeasurable. I dedicate this dissertation to you.

To Dr. Sashank Santhanam, your guidance and help in various stages has supported the completion of this work. I am extremely honored and grateful to have had the chance to work with you and learn from your work during the course of our PhD journeys. Thank you for being an invaluable mentor anna.

To my colleagues Erfan Al-Hossami and Zhuo Cheng, our journeys together as we worked towards our PhDs has been a source of support and encouragement which have made this work possible. Thank you for being the best colleagues I could have asked for.

To my friends, especially Ariana and Anusha, your presence has brought joy, laughter, and a sense of community to my life, reminding me of the importance of balance and camaraderie during the challenging times. I am fortunate to have you all in my life. Each interaction, conversation, and piece of advice has contributed to my growth, both personally and professionally. Your presence in my life has made a lasting impact, and I am honored to dedicate this dissertation to all of you.

To all the individuals who have played a significant role in my academic journey, including mentors, colleagues, and family members, I extend my heartfelt gratitude. Each interaction, conversation, and piece of advice has contributed to my growth, both personally and professionally. Your presence in my life has made a lasting impact, and I am honored to dedicate this dissertation to all of you.

Finally, to the pursuit of knowledge itself, I dedicate this dissertation. May it contribute, in its own small way, to the betterment of society and the advancement of human understanding. May it serve as a testament to the power of curiosity, perseverance, and collaboration.

Thank you all from the bottom of my heart.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Dr. Samira Shaikh, for her exceptional guidance and unwavering support throughout this dissertation journey. Her invaluable expertise, insightful feedback, and encouragement have been instrumental in shaping this work.

I would like to express gratitude to my dissertation committee members, Dr. Wlodek Zadrozny, Dr. Razvan Bunescu, Dr. Xi (Sunshine) Niu, and Dr. Mohamed Shehab, for their support and guidance throughout my PhD. Their feedback in all stages of my PhD journey has made this work that much better.

I would like to thank Dr. Sara Levens, Dr. Tiffany Gallicano, Dr. Cherie Maestas, and Dr. Wlodek Zadrozny, and my colleagues from the Social Media in Society grant team, for their guidance throughout the project, and beyond.

I would like to thank Dr. Sashank Santhanam for the support and guidance throughout most of my work.

I would like to thank Dr. Jing Yang for teaching me how to teach students during my assistantship.

I would like to thank the UNC Charlotte Graduate School for supporting my academic journey via the Graduate Assistant Support Plan (GASP), and the Graduate School Summer Fellowship (GSSF) programs.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

HCI Human Computer Interaction

LLM Large Language Model

MPC Multi Party Conversation

NLG Natural Language Generation

NLP Natural Language Processing

NLU Natural Language Understanding

PLM Pretrained Language Model

CHAPTER 1: Introduction

1.1 Background

Since the invention of artificial neural networks (circa 1940s), the goal of being able to converse naturally with computers has been a goal researchers have tirelessly worked towards. Within the field of Computer Science, Cognitive Science, and Computational Linguistics, technology has come a long way in being able to recognize patterns within natural language for various tasks, such as classifying the sentiment within a written sentence and being able to identify major topics which the conversations revolve around. Within the field of Computer Science, the sub-field of Natural Language Processing (NLP) focuses on approaches relating to understanding natural language (Natural Language Understanding, or NLU) and generating natural language (Natural Language Generation, or NLG). However, even with great advances in these research areas within NLP, the understanding and generation of multi-party communication (or group conversations) is relatively (and as we discuss in further sections, severely) understudied.

Two-party dialogue models, which model conversations between 2 participants, have seen great progress within the past 5 years, as evidenced by the frequency of surveys being written to capture the progress within NLG [17, 18, 19, 20, 21, 22, 23, 24, 25]. Approaches towards two-party dialogue modeling have ranged from rule-based systems to statistical approaches, to deep learning models with neural networks. Within the last decade, the field has developed a lot with the advent of pre-trained models which allow transfer learning for downstream tasks, introduced by [26, 27, 28], with the latest research bringing the field closer to a conversational AI system that can generate natural language text [29, 30]. While there are still open problems

that require further research, such as modeling long-term dialogue context modeling and infusion of knowledge, persona, and empathy [31, 32, 33, 25], dialogue system research has focused mainly on dyadic conversations, and there is a need for research in group conversation settings since they are just as common and relevant as dyadic conversations, called multi-party conversations (MPC).

With the increase in participants, group conversations pose several interesting challenges in addition to the challenge of generating natural language itself, the major ones including but not limited to:

1. **Speaker Identification**: understanding or predicting who the next speaker will be, as well as managing multiple speakers within the same conversation

2. **Addressee Recognition**: choosing the addressee(s) within the conversation while keeping in mind the roles of all other participants which could range from listeners to overhearers to eavesdroppers, as well as the implications of their roles

3. **Response Selection/Generation**: selecting a response from possible candidates or generating a response from scratch keeping the nature of the interactions between participants in mind

4. **Thread/Conversation Management**: managing different threads since a single group conversation could involve multiple subsets of participants talking about differing topics

Existing work tackling these tasks, whether separately or jointly, has focused on utilizing rule-based, statistical, and neural network based machine learning approaches, as is prevalent in two-party dialogue modeling. However, unlike the recent boom in the use of large language models (LLMs) and pre-trained language models (PLMs) in two-party dialogue modeling, multi-party dialogue modeling has just very recently evolved towards utilizing LLMs and PLMs [12, 16].

1.2 Motivation

An analysis of existing progress in the field of multi-party dialogue modeling shows how the field requires unifying, and highlights the various gaps within the field. My motivations for pursuing research in this area are listed below:

1. There is a lack of resources that are focused on multi-party conversations (MPC), especially more recent corpora which reflect the various changes in communication forms that have happened recently (more remote communication owing to COVID-19). We present a tool, which provides similar affordances as Twitter, that can be used for collecting data from informed participants. We utilize the tool to collect data and aim to release it publicly to contribute to MPC resources.

2. Many modeling systems recently presented for MPC do not utilize LLMs for generating utterances, although these have been shown to significantly boost the abilities of two-party conversations. We aim to utilize these towards persona-aware response generation, which has been shown to improve the quality of two-party conversational AI greatly and would be an important factor for a model to consider when multiple participants are present in a conversation.

3. A significantly fragmented piece in multi-party conversation modeling is how to evaluate these systems. As discussed before, significant challenges are introduced in modeling conversation between multiple participants, and thus the last part of my dissertation focuses on the need of the hour and learning from the confusion which has riddled two-party dialogue models for quite some time. We present an analysis of the evaluation methods used in existing research and how they could be improved.

1.3 Contributions

Here is a summary of high-level contributions made in this thesis:

1. We present a comprehensive survey of the field of multi-party conversation modeling. We classify existing corpora and systems into taxonomies which assist the synthesis of progress in this field in Chapter 2.

2. We create a mock social media platform which allows more straightforward collection of multi-party corpora in Section 3. We utilize this platform to collect and create a rich dataset of multi-party conversations along with speaker and addressee information for each utterance. We also collect persona-level attributes from participants and contribute this dataset for further study.

3. We present a novel response generation model towards persona-aware multi-party response generation. This helps the MPC response generation model speaker and addressee persona information, producing better responses. We demonstrate this work in Section 4.

4. We survey existing multi-party conversational AI work towards understanding inconsistencies and gaps towards evaluation. We report these inconsistencies, and expand on two-party conversation evaluation towards multi-party conversation evaluation in Section 5.

1.4 Outline

We present an overview of existing work and approaches taken toward multi-party dialogue modeling. Chapter 2 focuses on two surveys - one which provides details of multi-party corpora, and another which surveys the approaches taken towards solving either one, more, or all of the sub-tasks within multi-party dialogue modeling. The second survey also showcases various real-life applications of multi-party conversation systems which motivate the need for research in this area. Together these surveys

provide a path toward multi-party conversation modeling, and inform research contributions in this dissertation.

In Chapter 3, we survey existing corpora for MPC, organizing them into a taxonomy and providing details that relate to properties important for MPC dialogue modeling. Owing to the lack of asynchronous informal corpora discovered via this effort, we design and build a completely customizable mock social media platform that allows for easier data collection. We describe the tool and the resulting corpora we have been able to collect and contribute towards data for MPC modeling.

In Chapter 4, we present our work towards building a persona-based response generation system. We motivate the need for systems to take into account not only the existence of multiple speakers but also their different personas and the resulting relationships they display as a part of the conversation. We utilize these features to propose an MPC system towards better response generation given the different personas present in the conversation.

In Chapter 5, we discuss the challenges for evaluation introduced by group conversation settings. We focus on the basic errors which require reporting towards standards in evaluating dialogue models on all levels introduced in Section 1.1. We note inconsistencies in error reporting in existing research and discuss guidelines with an expanded integrated taxonomy that could help guide future benchmarks.

CHAPTER 2: Related Work

2.1 Introduction

The rise in neural network based approaches towards learning from big data has revolutionized the abilities of computers, achieving state-of-the-art performance on various tasks in the fields of computer vision, speech recognition, and various NLP tasks [34]. A major area within NLP that has seen progress with these techniques is NLG, with explosive growth in recent times as discussed in recent surveys [20, 35, 36]. However, most work in NLG focuses on two-party dialogue modeling, and while this focus is important owing to the open research problems such as modeling long-term dialogue context modeling and infusion of knowledge, persona, and empathy [31, 32, 33]; there is a pressing need to focus on group conversations which are just as, if not more, prevalent as dyadic conversations. Multi-party conversational agents stand to advance the future of work, finding applications in a huge array of situations - they can be integrated into formal settings such as meetings with teams, healthcare, search and rescue, as well as informal settings such as game-play. Particularly, with conversational home assistants such as Amazon Alexa, there is a push to develop AI to understand multiple users and act together as a team [37, 38].

Multi-party conversational systems have been shown to enrich the quality of interactions with humans in various domains, such as game-playing, meeting assistants, health assistants, and more. Some interesting applications exist in the health domain, where the ability of handling group conversations can allow family members and people with important information about the patient to assist more cohesively in patient care [39]. Another interesting example is the multimodal interaction capable robotic head called Furhat [40, 41], which has over 180 publications connected

to various research applications in group conversations such as meeting facilitation, couple counseling services, and educational assistance.

The approaches taken towards modeling multi-party conversations in existing work follow traditional approaches (involving using rule-based systems and state automata) as well as more recently proposed deep learning based approaches (using neural networks). These approaches aim to not only generate responses akin to two-party dialogue systems (such as probabilistic next word generation) but to also model speaker information and addressee recognition, along with the task of disentangling multiple threads pursued by sub-groups within the same conversation. Thus, the most studied sub-tasks within MPC modeling consist of Speaker Identification (SI), Addressee Recognition (AR), Response Selection or Generation (RS/RG), and Thread Management (TM). In this chapter, we survey over 330+ papers (bibliography released on Github [42]) and provide an overview of the challenges faced in all aspects (Section 2.2), available corpora and which tasks they have been used towards in past research (Section 2.3), comparisons to two-party dialogue modeling (Section 2.4), past and current approaches towards modeling multi-party dialogue, along with the formalized sub-tasks they aim to tackle (Section 2.5).

2.2 Challenges for Modeling Multi Party Conversations

The presence of multiple participants introduces new and interesting challenges for dialogue systems. If a Conversational AI system wishes to participate in a two-party dialogue, it is somewhat natural and straightforward for it to take on the role of the respondent when conversing with a user (Figure 2.1). Turn-taking is often consequential of the user's utterance, although in recent times there have been forays into challenges with handling multi-turn dialogue (multiple user inputs before bot response) as is often exhibited by everyday chit-chat [43, 44]. Response selection/generation also depends on the information provided or asked for by the user, and the main challenge here is to model the context of the conversation.

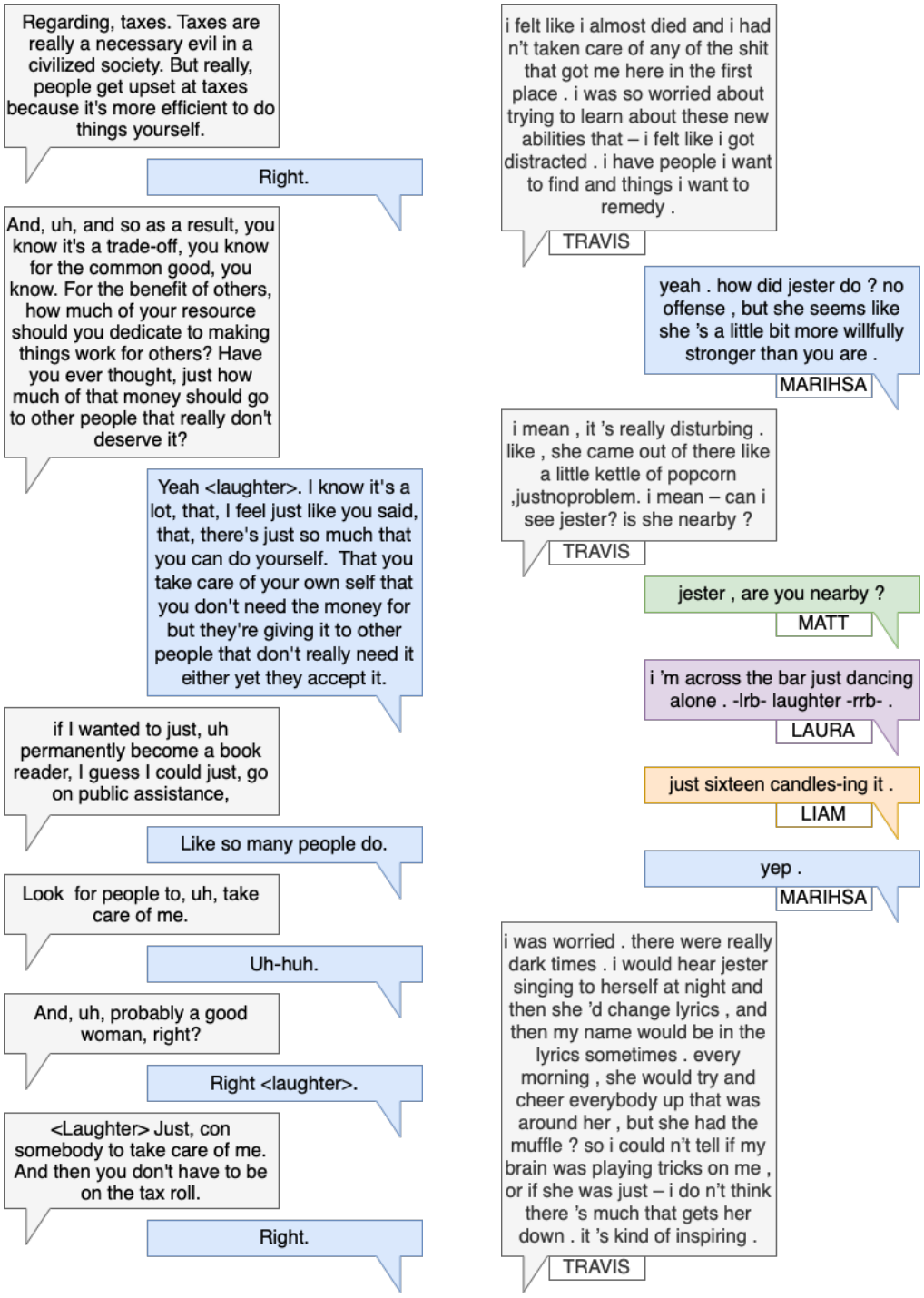


Figure 2.1: Two-party example taken from the Switchboard corpus [3]. Multi-party example taken from the CRD3 paper example [4]. Examples of informal unscripted spoken conversations - note the differences in the possibility of having different addressees (which could be multiple) in group conversations, as well as the possibility of having synchronously ongoing threads with sub-groups within the same conversation and group.

In contrast, a multi-party system does not have a straightforward way into the conversation. Turn-taking varies with the intents of all the participants within the conversation (speaker identification), and response selection/generation requires knowing which speaker said what (i.e. what to respond to and how to do so given conversation history). These facets make up the "interesting" challenges posed by multi-party conversations in direct comparison to two-party dialogue (Figure 2.1). Additionally, there are many "new" challenges posed as well, the most striking of which is the need to know the addressee(s) when taking a turn (addressee recognition). This includes the properties which might have been present for different participants, such as the designation of a participant as an eavesdropper or overhearer [45]. There is also the question of what kind of participation is expected from the system, one of three main types - (1) it could theoretically be an additional participant in the conversation (like note-taking for formal meetings, or (2) aim to emulate one of the participants already in the conversation (such as learning to participate as a character within a movie, or participating as "Laura" in 2.1), or (3) silently observe and model conversation flow for note-taking or discourse summarization. In the following discussion we present research which has attempted to formalize these challenges for multi-party conversations, and present these together in Figure 2.2 for easier reference throughout this chapter. We find that multiple avenues have been pursued: (1) with systems being additional participants in scenarios such as meetings, (2) with systems acting in place of one or more participants, and (3) with systems purely working on understanding conversational structure. We list the multitude of challenges faced by all three kinds of modeling here.

Through observations of similar interactions to CRD3 [4] between a robot in a situated virtual text-based environment, theoretical challenges [5] were listed according to the 6 W's (or 5 W's and 1 H) of information gathering, presented in Figure 2.2. These tasks were further formalized [6] specifically for MPC from a Human-Computer

Interaction (HCI) point of view: 1) *Initiative* - determining the frequency and type of each user's interactions with the system and how they affect the conversational flow; 2) *Dialogue Modeling* - to account for multiple threads of conversation, introducing new challenges for dialogue management and the subsequent need for joint or multiple state tracking; 3) *Error Handling* - error definitions would depend on the point of view of the participants (a response might make sense to one participant but not to another); 4) *User Modeling* - account for the differing participant personas; and 5) *Flexible "Multi" I/O* - dealing with cross-talk which is a natural part of multi-party conversation.

The aforementioned challenges are detailed further from a dialogue systems point of view [7] by observing the Mission Rehearsal Exercise (MRE) corpus [46] for dialogue modeling. The *Participant Roles* are broken down into a) *Conversational Roles* which includes two sub-issues - *who* can receive an utterance, and *who* it is addressed to; b) *Speaker Identification* which includes understanding *who* the speaker is so as to attribute the speech to the right participant; c) *Addressee Recognition* since the differentiation between listener and addressee becomes important in multi-party dialogue; and d) *Other Participant Roles* such as including support for specific behavior such as negotiation and an understanding of social roles. Following this is *Interaction Management*, consisting of a) *Turn Management* to understand *when* to take turn, release turn or keep turn; b) *Channel Management* which is applicable when multiple modalities are expected; c) *Thread Management* where multiple conversation threads co-occur with different participants and thus require management for each thread; d) *Initiative Management* to understand when to initiate new topics into the conversation, and when to keep the topic going; and e) *Attention Management* which consists of understanding when to add a new participant to the conversation and which thread each participant is paying attention to. Lastly, they include *Grounding and Obligations* in their discussion, since adding common ground to a conversation

might help bring the participants together or keep them true to the task at hand, and an understanding of the obligation of response is required at multiple levels - personally with each participant, subgroups that are discussing the same thread, and the group as a whole. This step is discussed further by presenting a *Question Under Discussion* (QUD) based model [47], which acts as a dialogue manager for addressing challenges related to grounding and obligation.

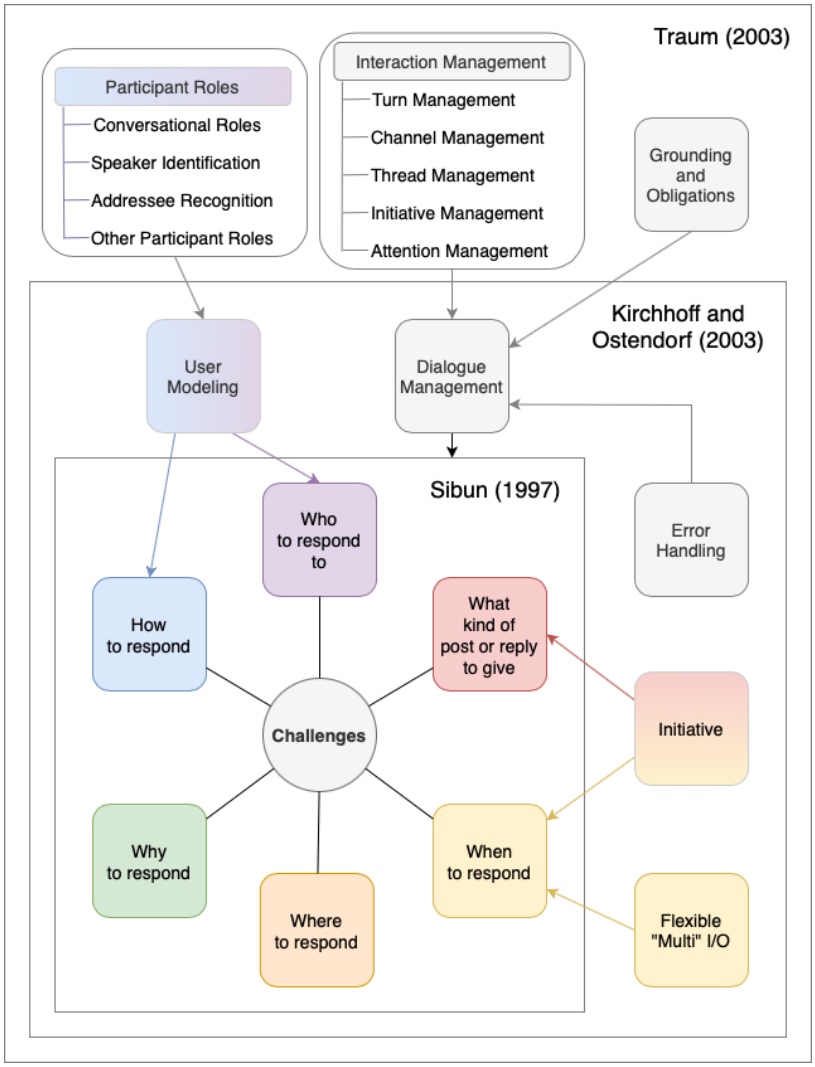


Figure 2.2: Challenges in multi-party conversations over two-party dialogue as presented in existing research [5, 6, 7].

While these articles discuss challenges in multi-party dialogue mainly from computational perspectives, some articles have approached these challenges from the

perspectives of the field of Psychology [48]. They argue that while two-party and multi-party conversations could be seen as qualitatively similar - common ground is accumulated in essentially the same way and participants exert the same types of influences upon each other - there are subjective factors that influence different types of conversations such as the overall goal of the dialogue, the overall number of participants, the type of dialogue, and more. Empirical observations studying personal home assistants [49] show how multiple participant interactions with them have nuanced issues such as having to repeat and refine queries, and the existence of mutual silences. They also find that members routinely organize their queries to the assistants either individually or collaboratively, showcasing challenges faced in industrial applications as well.

In the upcoming sections, we discuss these challenges with regards to the participation types discussed before - (1) active-additive, (2) active-simulative, and (3) passive - as seen in existing research. We also discuss future directions based on the research conducted within these roles.

2.3 Existing Multi Party Corpora

This section presents available corpora and their properties which have been utilized in multi-party conversation modeling in the past. The corpora here (a) have already been used in existing research in conversational systems; (b) have a text component, and focus on the English language; and (c) which include *multiple speakers in the majority of conversations.*

All included corpora are organized in a taxonomy (Figure 2.3), first categorized by whether they include *Spoken* or *Written* dialogue. Spoken corpora are further divided as *unscripted* vs. *scripted.* Within these type-based divisions, the corpora are then arranged by their main sources. The *unscripted spoken corpora* are thus arranged into 4 main categories - *informal discourse* mainly consisting of informal interactions such as radio talk shows, *formal discourse* mainly consisting of formal interactions such

as debates, *spontaneous speech* mainly consisting of spontaneous interactions such as teenage talk, and *meetings and interviews* mainly focused on data from sources such as TV interviews. Similarly, the *scripted spoken corpora* are arranged into scripts and dialogues from *plays*, *movies* and *TV series*. Lastly, the *written corpora* are arranged into four categories - *synchronous* mainly consisting of *chatroom talk*, and *online game-playing forums* with users mainly conversing about game progression; and *asynchronous* mainly consisting of posts made on online *forums* and short text messages on *microblog* websites with character limits for posts.

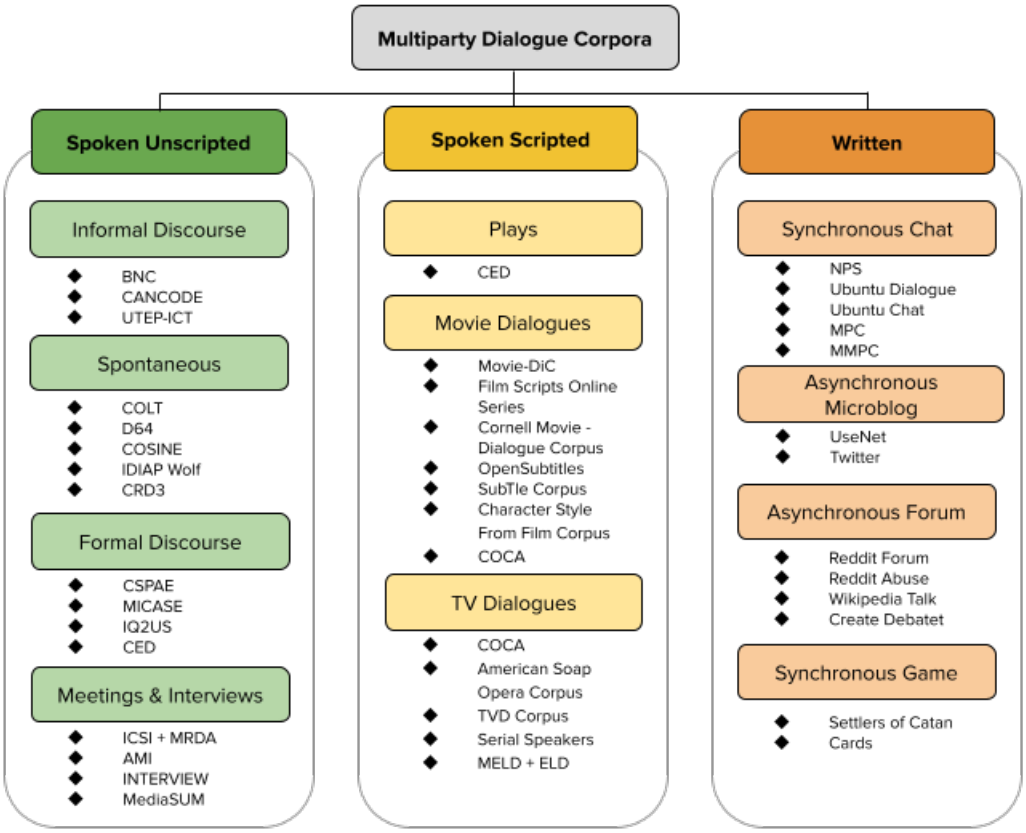


Figure 2.3: Taxonomy of available Multi-party Corpora, organized by source types - spoken unscripted, spoken scripted, and written.

Tables 2.2, 2.1 and 2.3 present additional details about each corpus, including the name and source citation, topics presented, quantitative details such as the number of dialogues, words, total length, and speakers, as well as whether they include other

modalities such as audio. The tables also include the *Task Descriptions* each corpus has been used for in the past, ranging from machine reading comprehension and turn-taking to speaker-identification.

Table 2.1: Details for all spoken, unscripted multi-party corpora. Starred (*) numbers are approximated from available information.

	Name	Topic	Num. dialogues	Num. words	Total Length	Total Speakers	Multi-modal?	Tasks
UNSCRIPTED INFORMAL	**Aggregated from various sources**							
	British National Corpus (BNC) [50]	Informal	854	10M	100 hrs*	23466	✓	word sense disambiguation, morphological & syntactic analysis
	CANCODE [51]	Informal	-	5M	550 hrs*	-	×	language learning, POS tagging
	Collected in specialized environments							
	D64 Corpus [52]	Natural	2	70K*	8 hrs	5	✓	involvement detection, studying silence and overlap in conversation
	COSINE [53]	Natural	10	160K	42 hrs	3.69 per session	✓	recognition of speech and speakers in noisy environments
	IDIAP Wolf Corpus [54]	Game	15	60K*	7 hrs	8-12 groups	✓	group performance in task-based interaction, implicit communication
	TEAMS corpus [55]	Game	116K	3M	47 hrs	3-4/ game	✓	entrainment, speaker transitions, personality identification & team dynamics
	Transcribed from pre-recorded media							
	COLT corpus [56]	Natural	100	500K	55 hrs	31	×	teenage talk trends
	CRD3 [4]	Game	159	5M	-	72	✓	character-action interactions in role playing games
UNSCRIPTED FORMAL	**Aggregated from various sources**							
	MICASE [57]	Academic	152	1.7M	200 hrs	1571	✓	male/female adjective use, academic discourse and vocabularies, English language learning
	Collected in specialized environments							
	AMI Meeting Corpus [58]	Formal	175	900K*	100 hrs	4-5 per meeting	✓	recognizing socio-economic roles, decision and action detection, summarization, dialogue act tagging
	ICSI MRDA [59, 60]	Meetings	75	795K	72 hrs	3-10 per meeting	✓	speaker overlap, summarization, speaker identification
	Transcribed from pre-recorded media							
	Intelligence Squared Debates [61]	Debates, predecided	108	1.8M	200 hrs*	3-5 per debate	✓	predictive models of debates, discourse modeling
	CSPAE [62]	Politics, education	200	2M	220 hrs*	400+	×	speech style & gender distinctions, variation between written & spoken corpora
	CED (1560-1760) [63]	Movies, formal	-	1.2 M	-	-	×	early English language variations and changes over time
	MediaSum [64]	Interview	463K	720M	-	6.5 per dialogue	✓	dialogue summarization
	INTERVIEW [65]	Interview	105K	126.7M	10K	184K	✓	follow-up question generation
	Canal9 [66]	Political Debates	70 debates	-	43 hrs	5 per debate	✓	speaker identification, turn-taking, conflict detection

2.3.1 Spoken Corpora

Spoken corpora is the most prevalent type of corpora available for MPCs. Spoken corpora presented in this chapter are further divided into two main categories

Table 2.2: Details for all spoken, scripted multi-party corpora. Starred (*) numbers are approximated from available information.

	Name	Topic	Num. dialogues	Num. words	Total Length	Total Speakers	Multi-modal?	Tasks
Transcribed from pre-recorded media								
SCRIPTED SPOKEN	Movie-DiC [67]	Movie dialogues	132K	6M	-	1-7 per dialogue	×	turn taking, speaker identification, emotional dialogue generation
	Cornell Movie Dialogue Corpus [68]	Movie dialogues	220K	9M	-	9035	×	
	Film scripts online series [69]	Movie scripts	263K	16M	1500 scripts	2-6 per script*	×	(information unavailable)
	OpenSubtitles [70]	Movie subtitles	337M	2.5G	-	2-6 per script*	×	turn taking, speaker identification, emotional dialogue generation
	SubTle corpus [71]	Movie subtitles	3.35M	20M	6184 movies	2-6 per script*	×	
	Character Style from Film Corpus [72]	Movie subtitles	151K	9.6M	862 movies	2-6 per script*	×	
	American Soap Opera Corpus [73]	TV dialogues	1.2M	100M	-	10-12 per script	×	
	TVD corpus [74]	TV dialogues	10K	600K	-	2-6 per script	✓	
	MELD [75]	TV dialogues	1400	109K	13.6 hrs*	400	✓	
	Serial Speakers [76]	TV dialogues	106K	682K	130 hrs	6 per script*	✓	
	MEISD [77]	TV dialogues	1000	50K unique	22 hrs	4072	✓	

(Tables 2.2 and 2.1) - *unscripted* which refers to spontaneous, unplanned dialogues; and *scripted* which refers to planned dialogue such as TV and movie scripts. The distinction between scripted and unscripted is made to allow for different modelling tasks, since scripted dialogue displays an absence of hesitations, repetitions and other normal non-fluency features.

2.3.1.1 Unscripted Spoken Corpora

One of the earliest multi-party spoken corpora is the British National Corpus (BNC) [50], originally created by the Oxford University press in 1980s-1990s. Covering a wide range of genres, including some written conversations, as well as POS-tagged data [93], it is important as a generalized multi-party conversation corpus. It has been used to study social differentiation in the use of English vocabulary [94], word frequency differences in spoken vs written text [95], and amplifiers such as "very" and "so" in the English language [96]. The Cambridge and Nottingham Corpus of Discourse in English (CANCODE) [51] focuses on interpersonal communication con-

Table 2.3: Details for all written multi-party corpora. Starred (*) numbers are approximated from available information.

Name	Topic	Num. dialogues	Num. words	Total Length	Total Speakers	Multi-modal?	Tasks
NPS Chat Corpus [78]	Informal chat	15	100M			×	part-of-speech tagging, dialogue act recognition
Ubuntu Dialogue Corpus [79]	Ubuntu OS Chatroom	930K	100M	-	-	×	speaker identification, discourse parsing, machine comprehension, response selection
Ubuntu Chat Corpus [80]	Ubuntu OS Chatroom	10655	2B	-	-	×	language learning, POS tagging
Molweni [81]	Ubuntu OS Chatroom	10K	24K	200 hrs	3.5 per dialogue	×	machine reading comprehension, discourse parsing
MPC Corpus [82]	Informal chatroom	14	58K	-	5 per session	×	turn-taking, speaker identification, detecting influence & leadership, group behavior
Settlers of Catan [83]	Informal, game-playing	21	-	-	2-6 players	×	modeling bargaining, negotiation, trading dialogue, risk-management in dialogue, action identification
Cards Corpus [84]	Informal, game-playing	1266	282K	-	-	×	goal-driven dialogue, event knowledge based questioning
Reddit Corpus [85]	Informal forum	84979	76M-414M*	-	521K	Maybe	discourse, cyberbully detection, exploring incel language
Reddit Domestic Abuse Corpus [86]	Abusive forum	21333	19M-303M	-		×	language biases, detecting harassment
Internet Argument Corpus [87]	Political forum	11000	73M	-	-	×	summarization, rhetoric and sarcasm, stance detection
Agreement in Wikipedia Talk Pages [88]	Informal	822	110K	-	-	×	linguistic tracing of manipulations, dialogue act recognition, social act recognition, conflict detection, speaker identification
Agreement by Create Debaters [89]	Informal	10000	1.4M	-	-	×	constructive disagreement, sarcasm, rumor classification, stance identification
Twitter Corpus [90]	Informal microblog	1.3M	125M	-	-	×	dialogue act recognition, author and topic identification, event discovery
UseNet Corpus [91, 92]	Informal microblog	47860	7B	-	-	×	modeling and analyzing text written on mobile devices

versations in various settings such as hair salons and restaurants. It has been used to study language use for teaching in classrooms [97], and is a resource for linguistic features of discourse. A more informal, casual English corpus is the Bergen Corpus of London Teenage Language (COLT) [56], which was recorded in secret to document spontaneous conversations and teenage language. It has been used to study trends in teenage language evolution [98], and is an excellent resource for spontaneous informal multi-party interaction.

The D64 Multimodal corpus [52] focuses on recording multi-modal dynamic interactions without specifying a topic, and has been utilized to study engagement in human-agent interaction [99]. The COnversational Speech In Noisy Environments (COSINE) [53] corpus introduces data collected in noisy environments, extending the challenges faced in multi-party dialogue such as turn-taking, and has been used to evaluate such systems [100].

The IDIAP Wolf corpus [54] focuses on group behavior in a competitive role-playing game setting, with a pre-condition of bad faith interactions similar to the "werewolf" or "mafia" game that makes it a unique corpus. It has been used in the AIWolfDial task to help train game-playing AI [101]. While specific instances of lying are not annotated, the "werewolf" of each game is annotated in the corpus. On the flip side, the TEAMS corpus [55] where teams of three or four speakers play two rounds of a cooperative board game, provides a novel resource for studying team entrainment and participation dominance. It has been used to build a novel graph-based vector representation of multi-party entrainment [102], gaining insights into the dynamics of the entrainment relations. The Critical Role Dungeons and Dragons Dataset (CRD3) [4] is a game-based corpus set in an open-ended scenario, with baselines on abstractive summarization benchmark and evaluation, based on each dialogue's summary.

Within formal settings, one of the oldest corpus is the Corpus of Spoken, Professional American-English (CSPAE) [62], consisting of two main components. The

first is White House press conferences, and the second is transcripts of meetings on national tests involving statements, discussions, and questions. In the past, it has proved a valuable resource for studying idioms and their usage [103]. The Michigan Corpus of Academic Spoken English (MICASE) [57] includes academic speech from university settings. It also comes with abstracts for each transcript, and has been used in online speech summarization [104]. Recent additions include data from interviews, such as the INTERVIEW [65] and MediaSum [64] corpora. They include transcripts from interviews on channels such as National Public Radio (NPR) [105] and CNN [106].

Debate-based settings are ideal candidates for multi-party corpora building, and thus the Intelligence Squared Debates (IQ2US) [61] are an important source. They follow an Oxford-style debating structure, and contain structured data making for a great resource for debate and argumentation analysis [107]. Canal9 [66] is another debate corpus, consisting of political debates. It includes a rich set of socially relevant annotations, and has been used in tasks such as conflict detection [108]. A historic debate corpus is the Trial Proceedings component of the Corpus of English Dialogues (CED) [63], which has been used to study signalling function in discourse [109].

Supplementing formal discourse in debate corpora are formal meeting corpora, with 2 corpora that have become really important for studying multi-party decision-making and discussions of actions to take are the ICSI meeting corpus [59], which also has Meeting Recorder Dialogue Act (MRDA) annotations [60]; and the multi-modal AMI meeting corpus [58]. ICSI has been used to further study multi-party language modeling [110], and AMI has been used to build summarization for meetings [111].

2.3.1.2 Scripted Spoken Corpora

Scripted spoken corpora consist of pre-defined scripts such as those for plays, movies, and TV series. These are inherently different as they are not spontaneous, and have pre-defined roles for speakers as well as information on when the dialogues

turns are taken. Some corpora are labelled with this information, while others are simply transcript-like (Table 2.2).

One of the earliest available scripted spoken corpora is a second component of the Corpus of English Dialogue (CED) [63] focusing on Prose Fiction. It has been used to study language styles in Shakespeare's plays in the context of contemporaneous plays [112].

The Movie-DiC Corpus [67] consists of a wide range of American movie scripts, along with context descriptions. It has even been used to generate parallel corpora for dialogue translation [113]. The Film Scripts Online Series [69] corpus includes British movie scripts, but is not available online. The Cornell Movie-Dialogue Corpus [68] contains metadata associated with each movie script, and has been used to generate emotionally aligned responses to dialogue [114]. The Character Style From Film Corpus [72] is another resource contributing towards guided text generation by providing character styles, created from the archive IMSDB [115]. It has been used to generate stylistic dialogue for narratives [116]. Both the OpenSubtitles [70] and SubTle corpus [71] are based on the OpenSubtitles website [117]. They are corpora of plain scripts, but the website continues to contribute as a resource for more data [118, 119].

Bridging the sources of movie and TV scripts is the Corpus of American Soap Operas [73] which focuses on informal language, and has been used to study cultural representation differences in American soap operas [120]. The TVD corpus [74] includes data from shows like The Big Bang Theory and Game of Thrones, supplemented by crowd-sourced contributions for tasks such as summarization. It has been used to build models for speaker identification [121]. The Serial Speakers [76] dataset supplements data from both the aforementioned TV serials by also including the House of Cards and additional annotations.

Recently, the Multimodal EmotionLines Dataset (MELD) [75] corpus has been

presented by extending the (ELD [122]), with audio-visual modality along with text. It has been used as a resource for Dialogue Act Classification [123]. The MEISD [77] dataset is built further with TV scripts from 10 series, adding Friends, How I Met Your Mother, The Office, House MD, Grey's Anatomy, Castle, Breaking Bad to the aforementioned series. FriendsPersona [124] focuses on annotated personalities of scripted characters based on the Big Five personality traits, consisting of 711 conversations from the TV show Friends. It was recently introduced, and has already been used towards personality detection tasks [125, 126].

2.3.2 Written Corpora

Written corpora for multi-party have often resulted from online chatroom discussions, like the NPS Chat Corpus [78], which is shared as a part of the NLTK [127], and is one of the first Computer-Mediated corpora. A chatroom corpus collected via mock messenger experiments is the Multi-Party Chat (MPC) Corpus [82] which presents an annotated corpus based on four levels with communication links, dialogue acts, local topics and meso-topics, and has been used to understand user roles and modeling leadership and influence [128].

The Ubuntu IRC chatroom [129] has also contributed significantly to corpora such as the Ubuntu Dialogue Corpus [79] and Ubuntu Chat Corpus [80], which were collected as users asked questions relating to Ubuntu on the forum, and other users answered them. They have been used to train many MPC systems [130, 8, 9, 12, 16]. The Molweni corpus [81] builds on the Ubuntu Chat Dialogue corpus, and adds annotations for machine reading comprehension and discourse parsing. Online forums such as Reddit [131], and Wikipedia [132] have also contributed to such corpora. These notably include the Reddit [85] corpus which has also been extended into larger corpora [133]. There have also been argumentative corpora obtained from online interactions, like the Reddit Domestic Abuse Corpus [86] taken from subreddits specific on domestic abuse, allowing for discourse analysis on this subject.

Game-playing corpora such as the Settlers of Catan Corpus [83] and Cards Corpus [84] are great informal additions to chatroom corpora, with a competitive environment albeit in an informal setting. They have been used for tasks such as training models for negotiation dialogues [134].

Debate and agreement corpora such as the Internet Argument Corpus [87], Agreement in Wikipedia Talk Pages [88] and Agreement by Create Debaters [89], from debate and discussion forums online such as CreateDebate [135] also contribute towards argumentation in dialogue research [136].

Additionally, there have been corpora obtained from social media such as UseNet [137] and Twitter [138]. These include the UseNet Corpus [91, 92], a platform which is considered a precursor to more recent forums; and the Twitter Corpus [90], which was intended to help model dialogue acts.

2.3.3 Special Mentions

This section includes special mentions of corpora as well as frameworks and toolkits that do not fall under previous categories.

There are very few corpora which have focused on human-machine conversations for multi-party interactions. The only such corpora existing to the best of our knowledge is the Mission Rehearsal Exercise (MRE) Corpus [46], which presents a dataset built as audio face-to-face sessions between human trainees and virtual agents. The main theme of the multimodal dataset is decision-making for a platoon-leader in a peace-keeping mission, with the trainee acting as a lieutenant. The corpora has about 30K words, 2K utterances, and a total of 55 speakers. Traum et al [139] also introduce another three-party negotiation dialogue corpus, called the Stabilization and Support Operations (SASO-EN) corpus, which grew out of experiments on the MRE corpus [140], focusing on eye-gaze behavior in three-party negotiation. In an example scenario, the data consists of a human user who plays the role of a captain whose mission is to move a local clinic to a safer location by negotiating with the doctor

and mayor of the city.

In the formal meeting and lecture space, the IDIAP meeting corpus [141] is another extension under the AMI project (AMI and ICSI were discussed in Section 2.3.1.1), which focuses on addressing behavior in multi-modal, multi-party, face-to-face conversations. The corpus additionally contains hand-annotated dialogue acts, adjacency pairs, addressees and gaze directions of meeting participants. The Computers in Human Interaction Loop (CHIL) is another corpus [142] which provides numerous synchronized audio and video streams of real lectures and meetings, captured in multiple recording sites over a period of 4 years, focusing on human interaction in smart rooms. Connected to formal spoken corpora, but focusing on the question-answering task in multi-party dialogue is the recently introduced QAConv corpus [143], with 34k questions taken from about 28k dialogues, with around 26k words and 32 speakers consisting of conversations taken from email, panels and other formal communication channels.

There are also several corpora, especially multimodal, which have been transcribed, but we could not find the statistics. These include the VACE multimodal meeting corpus [144], which investigates the interaction among speech, gesture, posture, and gaze in meetings. Another corpus is the MULTISIMO corpus [145], towards modeling of collaborative aspects of multimodal behavior in groups that perform simple tasks between two people, supported by a facilitator. Mana et al [146] also present the Mission Survival Corpora (MSC) 1 and 2, a multi-modal corpus of multi-party meetings, automatically annotated using audio-visual cues (speech rate, pitch and energy, head orientation, hand and body fidgeting). Due to the limited information available, we do not add these corpora to the tables or the taxonomy.

A variation of the Machines Talking to Machines framework [147] allows a simulated user bot and a domain-agnostic system bot to converse to exhaustively generate dialogue "outlines", i.e. sequences of template utterances and their semantic parses,

which can then be contextually rewritten by crowdworkers to maintain saliency and coherence while preserving meaning. ChatArena [148] is another such framework which allows studying multi-agent interactions in game environments. We include these frameworks as they could contribute to collecting data for multi-party dialogue by extending it to include more simulated users and bots.

We also make special mention of the Convokit tool [85], which is a toolkit for downloading corpora for dialogues. It allows the downloads to follow standard format for all available corpora. It also provides the functionality to load custom datasets in a similar format, making it easier to work with multiple corpora at once.

2.3.4 Data Collection Methods

Several methods of data collection have been used to collect the aforementioned corpora. We organize these into three main categories and discuss in detail below.

Aggregated from various sources. BNC, CANCODE, and MICASE employ the aggregation method to build the corpora. They pull information from various sources, including text from sources such as newspapers, journals, publicly available government meetings, radio phone-ins, academic writings, seminars, advising sessions etc. These corpora incorporate multiple types of speech, and often include speech surrounding multiple topics (especially BNC and CANCODE, MICASE mainly focuses on academic settings to collect data). They are thus great candidates for studying language semantics and have been employed to study large-scale vocabularies [149] and word sense disambiguation [150] in the past.

Transcribed from pre-recorded media. Single (or double) source origins, such as COLT, CRD3, and IQ2, maintain focus on certain themes, such as formal meeting data. These are not collected within specialized environments, but consist of either transcribed speech recorded in the wild, transcribed interviews & meetings, and online forum or social media data. This category also includes scripted corpora, which are usually collections of various scripts & dialogues from plays, movies and TV series,

such as TVD and SubTle. Having a set theme allows these corpora to be used for generating themed text such as MELD being used for character identification as a part of the 2018 SemEval challenge [151].

Collected in specialized environments. Most multi-modal corpora employ specialized environments or equipment to collect data that can be synchronized across multiple modalities. Most focus on data collection using *audio*, which can then be transcribed. Specialized room environments with studio-quality recording (ICSI, AMI), close-talking mics (ICSI, IDIAP Wolf, TEAMS), and a combination of far- and close-field mics (COSINE, AMI) have provided better data collection for corpora, allowing for annotations of speech activity and pauses as well. Another popular data collection method focuses on *video*, such as motion sensing (D64), and video cams (IDIAP Wolf, TEAMS, AMI), which supplement speech data well by also allowing for annotation of head movement, gesture, and eye-gaze tracking. There are also multiple projects that emulate online social media platforms for controlled data collection, such as the Truman platform [152] and Community Connect [153].

With these available corpora and their properties, we now move the discussion back toward MPC modeling, beginning with how research has studied similarities and differences between two-party and multi-party conversations.

2.4 Comparisons to Two Party Dialogue Research

This section discusses research which has tried to empirically observe or experimentally discover the challenges introduced by multiple participants and their behaviors in multi-party conversations.

2.4.1 Direct comparisons of two party and multi party analysis

Perhaps one of the earliest and direct analysis of multi-party conversation challenges in comparison to two-party dialogue, especially towards **clausal ellipsis** (omissions of clauses), was conducted by Ricento [154]. They argue that the inherent structure of

a discourse unit changes when a conversation takes place in multi-party settings due to multiple participants sharing thoughts together, and as a result, if a discourse unit was analyzed in isolation, it would appear quite different in two-party vs multi-party conversational settings. They present observations owing to the sharing of consciousness, competition for the floor, and on-going joint construction of discourse, relating to the challenges presented by Traum [7] - specifically Interaction Management and Grounding and Obligations (Figure 2.2). They theorize that these behaviors could be predictable, owing to observations regarding topic control - once a topic is introduced and established in discourse, cooperative participants attend to it. They also observe that with conversations involving more than 2 actively participating speakers, clausal ellipses tend to be more common. Ishizaki and Kato [155] limit this analysis to three-party conversational analysis as compared to two-party, finding that even one more participant changes the conversational characteristics. They find that while there are similar properties, three-person conversations show more **initiative-taking** behavior in their empirical study (relating to the challenges of Initiative in Figure 2.2). Lastly, Yoon and Brown-Shmidt [156] study **audience design** from a participant point of view in multi-party conversations - they test whether speakers encode the perspective of multiple addressees, and then simultaneously consider their knowledge and physical context during referential design in a three-party conversation, and find that they do take into consideration both addressee knowledge and physical context when designing utterances, consistent with a *knowledge-scene integration view*, relating to multiple challenges from Figure 2.2.

2.4.2 Discourse challenges in multi party conversation modeling

Research towards discourse in MPC has ranged from modeling turn-taking to context modeling for negotiation and co-ordination or collaboration, in comparison to two-party conversations.

Aoki et al [157] analyze where a system should pay attention to study how partic-

ipants initialize and carry conversations, or **turn-taking** (Interaction Management in Figure 2.2), in multi-party interactions. They present an analytic study focusing on quantifying *participation sequences*, which they define as participant turns that are related to adjacent turns in the sequence. Their findings motivate an update to their data collection procedure - their original prototype extracts features from a relatively short time window, considers only turn-taking features, and identifies floors through analysis of (aggregated) pairwise measures of turn-taking. Based on their findings from the study, they propose re-engineering the system around a segment-based architecture, expecting the updates to provide a principled and more straightforward framework to consider data over longer time spans, to include features other than those based on turn-taking (such as potential schisming-related events), and to analyze behaviors that span groups (such as coordinated actions). Howes et al [158] present a study of speaker contributions for **turn-taking** as well as **negotiation** based on artificially introducing contributions in conversation which include utterances that either continue or complete earlier utterances, called *compound contributions* (CCs). Their motivations arise from the observation that nearly one fifth of all contributions in naturally occurring dialogue continue some previous contribution. They find that participants are able to process and interpret fake CCs successfully, despite their arbitrary split points (points where a CC splits into antecedent vs continuing information with respect to the speaker of the antecedent vs continuing information) and not being explicitly aware of the experimental manipulation. They also find that the utterances of the participants are affected by expectations of who will speak next and whether there are subgroups which agree or disagree on ongoing topics. From a computational modelling point of view, they find that start-completeness of continuations is rare, and that a dialogue system may have a chance of detecting continuations from surface characteristics of the input, however, there are no restrictions on where a split point may occur and this might be difficult to

model.

McBurney, Hitchcock, and Parsons [159] present a foundational study focusing on mathematically formalizing deliberation, or **negotiation discourse**, in multi-party conversations specifically between multiple autonomous software agents, presenting a formal model and applying it in game-playing conversations. Deliberation differs from negotiation in that the resource being negotiated upon is scarce, or time-sensitive. They present the properties and stages involved in such dialogue following previously established rules for discourse [160], the stages and corresponding agent actions being: 1) *Open* - `open_dialogue` or `enter_dialogue`, 2) *Inform* - `propose, assert, retract` or `ask_justify`, 3) *Propose* - `propose`, 4) *Consider* - `assert, prefer` or `ask_justify`, 5) *Revise* - `propose`, 6) *Recommend* - `move, assert` or `reject`, 7) *Confirm* - `move` or `assert`, and 8) *Close* - `withdraw_dialogue`. They call this framework the Deliberation Dialogue Framework (DDF) Protocol, noting that their grounding of the framework in an argumentation-theoretic account of deliberative decision-making means that the framework's sentence types, dialogue stages and locutions are specific to deliberation dialogues, and that in settings where participants might also have objectives other than sincere deliberations, DDF might not be able to express and code their various objectives. However, they find that this might be true of all existing computational frameworks.

Healey and Mills [161] study semantic **co-ordination** arguing that participants with different degrees of involvement in an interaction develop different levels of communicative co-ordination with one another. They compare two previously suggested models and apply them to multi-party conversations captured for game-playing while solving a maze - the grounding model [162] and the interactive alignment model [163]. These models attribute differences in semantic coordination achieved in multi-party exchanges to differences in participant's opportunities for interaction. The grounding model predicts that speakers actively track the different levels of co-ordination that

develop with different participants, whereas the interactive alignment model predicts that speakers respond instead to the cumulative exposure to particular inputs independently of their origin in the conversation. However, they find that neither of these models provide an accurate account of co-ordination in their multiple participant maze-solving game, drawing attention back to the Grounding and Interaction challenges (Figure 2.2). Eshghi and Healey [164] study dialogue contexts in multi-party conversation surrounding **collaborative task-solving** based on tangram matching, involving two directors who can see half the tangrams which need to be arranged in order and 1 matcher who can see all tangrams and match them with directions from the directors. They find that even within small groups of fully ratified conversational participants distinct conversational contexts can emerge as a result of changes in the primary participants, and hence they are, as indexed by ellipses, anaphora and other context sensitive expressions, more fine-grained than interactional units such as an F-formation [165]. As a result, they argue that there is a need to index these contexts to specific sets of participants and provided conditions, in terms of the participants' statuses during a stretch of talk/sub-dialogue. They also argue that attempts to scale-up computational/formal models of dyadic to multi-party dialogue need to take these findings into account, since they find that there are changes with respect to how to track and update content as the conversation moves forward and what is salient for whom at any given point.

Strzalkowski et al [166] present a study focusing on a two-tiered approach which first detects and classifies social cues such as topic control and disagreement, which in turn help observe higher order behaviors such as **leadership** (User Modeling and Initiative in Figure 2.2), using the MPC corpus [82]. These are important factors to consider owing to their emergence specifically in group conversations. Their presented model, called Detecting Social Actions and Roles in Multi-party Dialogue model (DSARMD-1), which labels *Communicative links, Dialogue Acts, Local top-*

ics, and *Topic reference polarity* based on English language annotations provided to it. Based on these labels along with supplementing automatic measures, they study *Agenda Control*, *Disagreement* and *Involvement* in dialogue. They find that this paradigm, although built specifically for their use-case, performs well (80% accuracy) on their corpus. They further extend their analysis to Mandarin Chinese [128], with 70% accuracy. Their research points towards future work, where considering social cues (and thus social roles) could be an important factor towards modeling agents that can participate in MPC more meaningfully.

2.5 Advances in Tasks Towards Multi Party Conversation Modeling

We break down challenges introduced by the presence of multiple participants, as discussed in Sections 2.2 and 2.4, into specific tasks which have been the focus of research in the past, and present them in this section. We organize this section by talking about existing research in Participant Roles and Dialogue Management, further discussing tasks relevant towards natural language understanding such as work in dialogue acts, anaphora resolution, entrainment and topic identification. Some approaches aim to model more than one of the tasks described above, and these are included in a subsection of their own. We also include which types of systems each task pertains to, and how it can contribute to systems of that type (namely (1) active-additive, (2) active-simulative, and (3) passive).

2.5.1 Participant Roles

Research in modeling participant roles includes addressee recognition (who to reply to) and speaker identification (who to listen to, or if speaking as a system, then which participant to model).

Early research in addressee recognition included statistical approaches towards figuring out markers such as interaction history, meeting action history, and user and spatial context among features from other modalities such as gaze and gestures [167],

although the effectiveness of each of these features are not empirically verified. Similarly, for modeling **speaker identification**, Hawes, Lin and Resnik [168] hypothesize that discourse markers and personal references provide important features for first- and second-order *Conditional Random Fields* (CRFs), for a corpus of court proceedings. They create a feature set involving unigrams, discourse markers, and personal references. They find that there are predictable patterns of interactions between justices in the oral arguments (as evidenced by the performance of the second-order models) and discourse markers along with personal references provide important clues for underlying discourse relations (at least those relevant in governing turn-taking behavior). Most early work thus shows a trend towards passive (type 3) systems for MPC through statistical methods.

More recent work, within the past decade, tends to utilize more neural networks based methods. Ma et al [169] are perhaps the first to use neural networks, as well as an active-simulative (type 2) system, to model speaker identification. They utilize *Convolutional Neural Networks* (CNNs) towards modeling conversations from the OPUS corpus [70]. They also apply several approaches and compare results - using 1) *K-Nearest Neighbors*, *Recurrent Neural Networks* (RNN), 2) Baseline CNN, 3) Multi-document CNN (which takes in all utterances that are part of the same scene as a single input + the pooling and prediction layers are fully connected), 4) CNN with surrounding utterance (where each utterance vector is concatenated with both the previous two utterances and the subsequent utterance in the same scene to preserve dialogue structure information since TV dialogues are highly structured), and 5) CNN with utterance concatenation (where all the utterances for each speaker in one scene are concatenated in the original dialogue order). They find that the Multi-document CNN performs best, with an accuracy of $\tilde{3}1\%$, and that considering neighboring utterances helps the models by providing context.

de Bayser et al [170] build on Ouchi and Tsuboi [8] (refer to Section 2.5.2), focusing

more so on finding whose turn to speak comes next by employing several different Machine Learning approaches as both active-additive (type 1) and active-simulative (type 2) settings, including *Maximum Likelihood Estimation* (MLE), *Support Vector Machine* (SVM), *Convolutional Neural Network* (CNN), and *Long Short-Term Memory network* (LSTM). They find that an agent-and-content CNN, which maintains agent encodings for each speaker and content encodings for each utterance by each speaker, performs best, albeit almost at the level of a very simple, baseline model (Repeat-Last) which is a simple rule for prediction. They further extend their work [171], building their system by including a hybrid of multiple approaches, particularly rule-based *finite state automata*, to achieve better performance on their task.

Qiu et al [172] propose a *Variational Recurrent Neural Network* (VRNN) [173], which incorporates a non-projective dependency tree attention layer to learn dialogue structure in an unsupervised manner. This is the first active-additive (type 1) system, which does not rely on previously known participant information to predict the speaker-addressee pair, and thus being able to handle being an additional participant into the MPC. They find that their proposed model is capable of distinguishing speakers and addresses by constructing an utterance dependency tree, automatically disentangling dialogues without explicit human annotation on the Ubuntu Chat corpus [80], with similar performance to GSN [174] (refer to Section 2.5.2).

The task of identifying participant roles thus includes research in all active-additive, active-simulative, and passive system types, with a few even focusing on more than one single system type. There is further scope for extending these abilities towards making MPC systems more capable of participating in various kinds of conversations, especially toward formal interactions.

2.5.2 Dialogue Management

Research in dialogue management includes response selection (given a few possible candidate responses, which one fits best) or generation (generating a response from

scratch based on who is speaking to whom). We first present research which has focused on methods for both response selection and addressee recognition from Section 2.5.1 to bring to fore the challenges faced by multi-party dialogue systems - research on tackling these facets separately is more recent, which might indicate that the individual tasks could be easier to tackle.

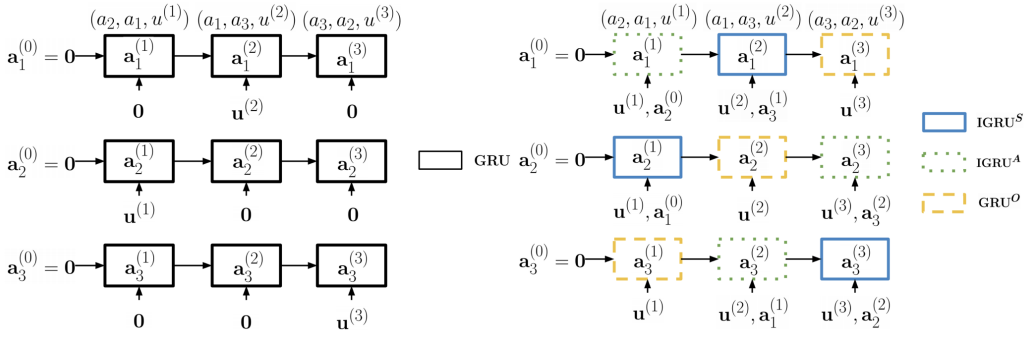
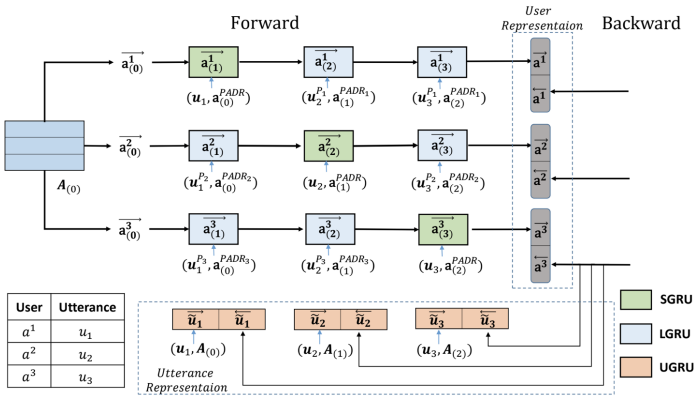


Figure 2.4: Visual representation for DynamicRNN [8] and SI-RNN [9]. Both model turn-taking along similar frameworks where the input is (responding agent, context, candidate responses) and the output is (addressee, response). While DynamicRNN can track the speaker status by capturing *who* says *what* in multi-party conversation it focuses only on updating sender's understanding for each utterance, whereas SI-RNN updates embeddings for all the speakers (based on whether they are the sender, addressee or observer) besides the sender at each time step, thus accounting for all participants during every utterance. Another key difference is that addressee-response pairs are separately modeled by DynamicRNN, whereas SI-RNN models them jointly.



User	Utterance
a^1	u_1
a^2	u_2
a^3	u_3

Figure 2.5: Who2Whom (W2W) system representation [10] - it models conversations similar to SI-RNN [9], however W2W aims to predict missing addressees as well, thus expanding the task. W2W additionally scans the conversation session from two directions and the representation of each user and utterance is the concatenation of both sides.

2.5.2.1 Response Selection and Addressee Recognition

Ouchi and Tsuboi [8] formalized the task of **addressee and response selection** jointly for multi-party conversation. They introduced a conversational system for tracking the complex interactions between a multiple speaker setup, which could jointly model who the speaker was and what they were talking about in context, proposing that the input include a context, a responding speaker, and a target addressee; and the output focus on both the response as well as the intended addressee. This system had two frameworks, one static and the other dynamic, of which the dynamic framework performed best. Thus an active-simulative (type 2) system, called *DynamicRNN*, solved the task in two phases: 1) speaker embeddings to track each speaker status, which dynamically changes with time step t, and 2) producing the context embedding from the speaker embeddings to select addressee and response based on embedding similarity among context, speaker, and utterance. The model is shown in Figure 2.4.

Building on this, Zhang et al [9] then proposed the *Speaker Interaction Recurrent Neural Network* (SI-RNN, presented in Figure 2.4), which tackled this problem while also taking into account the phenomenon of role-changing in the natural flow of multi-party conversations, outperforming DynamicRNN. Another active-additive (type 2) system, they define role-changing as the change of role between speaker, addressee, and observer. In Figure 2.4, it is possible to observe these role transitions for the speakers (and related speaker embeddings) as the utterances in the dialogue progress. SI-RNN was also patented in 2020 [175]. Le et al [10] then proposed learning not only the last addressee and their role, but also every missing addressee in the conversation with their *WhoToWhom* system (W2W, presented in Figure 2.5) during the entire session. The proposed W2W model had three main steps - 1) initialization of utterance and user representations, 2) interactive representation learning of users and utterances, and 3) matching procedure for identifying the addressee. While

initializing, they also encode spatial information about who the predecessor and successor speakers are by sorting the user representation in a descending order according to the first time when they speak. This emulates the probability of a person speaking replying to the person who spoke before them or addressing the person who spoke after them. The interactive representation learning extends the proposed system in SI-RNN [9], now including a *Person Attention Mechanism* to track state progress. The matching procedure includes fusing the information learnt by the utterance embedding, current speaker embedding and the user-summary vector into an enhanced utterance embedding. The similarity between this fused embedding and each listener determines the addressee. This compounded approach improves upon the previous baseline significantly.

2.5.2.2 Response Selection

The Thread-Encoder model [176] incorporates dialogue dependency information into the response selection task, demonstrating that dependency relations in the dialogue history are useful in predicting dialogue responses. They design an algorithm to extract several threads from the dialogue history as an active-simulative (type 2) system. The architecture of the Thread-Encoder model consists of two main layers: 1) *Encoding* layer using two *Transformers* for thread encoding (all of the turns in a thread are concatenated into a long sequence in reverse chronological order as the input) and candidate encoding, and 2) *Matching* layer which uses an *attention* layer to distill information from threads by attending the threads to the candidates. They take in each dialogue as a triplet of context, response, and whether or not that response was the correct one to train the proposed model. They find that their approach outperforms previous approaches, claiming the spot for the new state-of-the-art (as of 2021).

Topic-BERT [11] frames response selection as a dynamic topic tracking task to match the topic between the response and relevant conversation context and propose

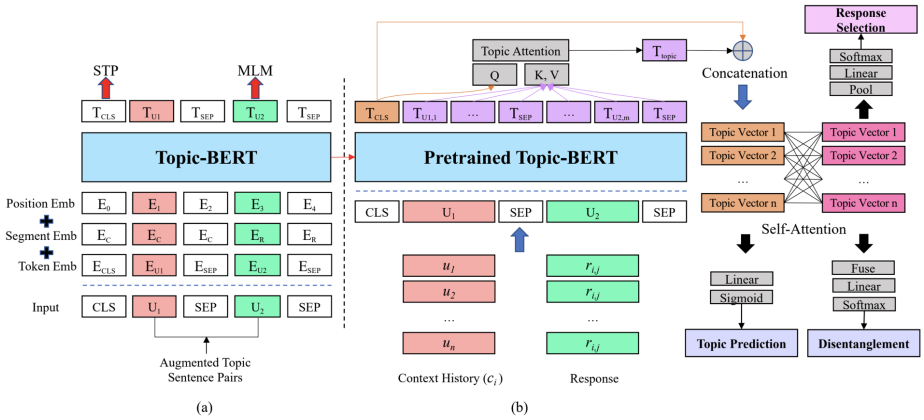


Figure 2.6: Overview of Topic-BERT structure [11] (a) Topic-BERT pre-training with topic sentence pairs to incorporate utterance-utterance topic relationship (b) Multi-task framework which uses the pretrained Topic-BERT to enhance topic information in the encoded representations to support three downstream tasks - response selection as the main task while topic prediction and disentanglement as two auxiliary tasks.

a novel multi-task learning framework, albeit based on dyadic interactions. The model encodes topic information which combines response selection task with topic prediction and topic disentanglement,, an active-simulative (type 2) system. The aim is to track how the conversation topics change from one utterance to another and use it for ranking the candidate responses. It encodes an utterance from the context, along with a candidate response, using a pre-trained Topic-BERT encoder. The contextual token representations in Topic-BERT encode topic relevance between the tokens of the utterance and the tokens of the candidate response, while the class representation captures utterance-level topic relevance (Figure 2.6). The authors find that their results also become state-of-the-art on the response selection task. However, comparing performance [176] is difficult since evaluations are not made on similar metrics, which is a known issue in this research area [177].

Speaker-Aware BERT (SA-BERT) [178] also utilizes BERT to build a speaker-aware response selection model which can handle speaker change information for the Ubuntu Dialogue corpus [79], an active-simulative (type 2) system. It tackles response selection with two sub-tasks, 1) *Speaker Embeddings & Segmentations* to

distinguish utterances in a context and model the speaker change in turn as the conversation progresses by utilizing speaker embeddings and segmentation tokens, and 2) *Speaker-Aware Disentanglement Strategy* for identifying the multiple speakers' roles in multi-party conversation. Using pre-trained BERT helps them achieve the current state-of-the-art performance for response selection. However, once again the performance achieved by these models is not directly comparable to peers - even with similar datasets being used - owing to differing evaluation metrics which are reported.

2.5.2.3 Response Generation

The tree-based group conversation model [179] (Figure 2.4) organizes the utterance flows in the conversation into a tree-based frame, designed especially for the group conversation scenario, based on experiments using the Ubuntu IRC Chat logs [80]. The *tree-based formulation* consists of constructing an initial tree purely from chat logs where each node is an utterance, then *splitting* any shared nodes into duplicate unshared nodes, and then using *hierarchical encoding* with *Gated Recurrent Units* (GRUs [180]) to create representations for encoding the inputs. The response is generated by using a *Decoder* which takes in the input - a context vector and the embedding of a previously decoded word to update its state using another GRU. Using this method for an active-simulative (type 2) system, they argue and find that while the method is rather simple, it indicates that less context information results in better generations in group conversations.

Interlocutor-aware Contexts into Recurrent Encoder-Decoder frameworks (ICRED) [181] focus instead on response generation in multi-party dialogue on a word-by-word basis to model speaker interactions. They change the previously used input-output format [8] using input consisting of (context, responding speaker, target addressee) and output being a generated response, another active-simulative (type 2) system. They introduce an end-to-end framework built using multiple layers - consisting of an *utterance encoder layer* which transforms input utterance into distributional rep-

resentations similar to those of SI-RNN [9]; a *speaker interaction layer* which utilizes the interactive speaker encoder similar to SI-RNN; an *addressee memory layer* which memorizes the contextual word representations in the last utterance said by the target addressee, and the contextual representation for each word obtained from the utterance encoder layer; and a *decoder layer* generates the output from the contextual speaker vector, contextual addressee vector, and attentional addressee vector. This approach allows the generated response to be novel, which is a limitation of SI-RNN, and the paper states that the results remarkably outperform strong baselines on automatic and manual evaluation metrics.

HeterMPC [16] introduce a graph-based approach towards response generation as an active-simulative (type 2) system, building further on their previous work [12]. Drawing from GSN [174], the paper changes the homogeneous nature of GSN into a heterogeneous graph with directed edges, allowing modeling of utterance and speaker information. With ablation experiments, they showcase that their heterogeneous graph approach outperforms GSN on the Ubuntu IRC [79] corpus with a statistically significant margin.

Research in Dialogue Management has thus mostly focused on active-simulative system types, which shows that there is scope for advancing research in active-additive and passive system types. These could involve systems which can help guide conversations or act as moderators in a conversations, and systems which focus on predicting speakers and recognize addresses. It is also imperative to have more diversity in the corpora used for modeling, since most research is limited to the Ubuntu IRC corpus [79, 80]. This points to a need for better data collection for MPC, a future direction for MPC research [182].

2.5.3 Dialogue Acts and Discourse

Dialogue acts are important features of a corpora for dialogue systems, allowing for a structured representation which facilitates the study of the tasks discussed so

far. Annotations for dialogue acts are often used as features when training dialogue systems to model dialogue. Multi-party conversations introduce new challenges as discussed before, which thus require the dialogue acts annotations and classification methods to account for additional properties. In this section, we discuss articles which have discussed these challenges, limiting discussion to previous work which has focused on multi-party settings. We start with a discussion surrounding data annotation and visualization methods for MPC, then standards for dialogue act annotations towards various tasks important for MPC modeling, and then further discuss methods to perform dialogue act classification based on these annotations.

2.5.3.1 Data Annotation and Visualization Methods

Building data-driven systems involves the use of corpora as well as the need to annotate them for modeling tasks. This section focuses on the tools which have been applied by this existing research towards multi-party conversations for annotation, transcription and visualization tasks.

Garg et al [1] present an analysis of various tools as they use them towards *transcribing* and *annotating* the MRE corpus [46], shown in Table 2.4. The authors weigh the factors in the table, giving priority to "ease of use" by the annotators and ease of import and export of data. We update the table to only include tools which have been maintained, noting how recent the last update was. To the best of our knowledge, these are till date the only released software for multi-party conversation annotation.

Apart from these tools for annotation in transcription and visualization for multi-party conversation, discourse *annotation guides* were introduced by Besser and Alexandersson [186] based on the AMI meeting corpus [58]. Popescu and Caelen [187] present a segmented discourse representation structure of the challenges involved with annotating and thus understanding dialogue acts such as turn-taking in multi-party discourse interactions, in keeping with Segmented Discourse Representation Theory (SDRT) [188]. They also describe an algorithm for the participation of an

Table 2.4: Evaluation of tools for annotation, transcription, and visualization for multi-party data [1]. Entries marked with a "+" indicate that the tool performed well in that category, and a "-" means it could have performed better. For non-binary evaluation metrics, the category was marked based on how well it fit the minimum expectations of the authors. We show only tools that have been updated in recent times, and add last updated information as of May 2023.

Evaluation	Praat [183]	Anvil [184]	Transcriber [185]
Portability	+	+	+
Source Code	+	?	+
A/V Interface	+	+	+
Comments	-	-	+
Coding Scheme Flexibility	N/A	+	N/A
Viewing Work	-	-	+
Ease of Use	-	-	+
Support/Manual	+	+	-
Overall	+	-	+
Last Updated	05/2023	08/2017	03/2017

agent in multi-party discourse based on the emergent structures. However, empirical evaluation of the proposed algorithm is mentioned as a part of future work. Li et al [189] provide a graph-based method to *annotate* the structure for discourse-level parsing and machine comprehension of Ubuntu Chat [80] and Settlers of Catan [83] corpora. They suggest annotations for 16 discourse relations - *Comment, Clarification question, Elaboration, Acknowledgement, Continuation, Explanation, Conditional, Question-answer pair, Alternation, Q-Elab, Result, Background, Narration, Correction, Parallel, Contrast* for each conversation.

Gilmartin and Campbell [190] present a *visualization* tool called STAVE while studying the d64 corpus [52] for discourse analysis. The software also generates colour-coded transcripts in Conversation Analysis format from a simple transcription file. They further update the tool to allow annotations for disfluencies for speech-level (interruption points, pauses, overlaps), word-level (unfinished, contracted, repeated, substituted, deleted) and utterance-level (unfinished) [191]. Furthermore, *ArgViz* [192] focuses on interactive *visualizations* to analyze dynamic topical structure (which are in turn determined by the SITS algorithm [193], (described in Section 2.5.6), while *TeMoCo* [194] focuses on temporal exploration specifically in healthcare clinical

settings.

2.5.3.2 Dialogue Act Annotation Taxonomy

Most dialogue act annotations which have been utilized in multi-party research have either been built based on multi-party conversation or modified from a pre-existing taxonomy to apply towards multi-party scenarios. Popescu-Belis [195] provide an in-depth survey comparing annotation tagsets from multiple dialogue act annotation taxonomies, including DAMSL [196], SWBD-DAMSL [197], ICSI-MRDA [59, 60] and MALTUS [198], of which the latter two are utilized in multi-party corpora annotation. They provide interesting insights into theoretical and practical applications of the annotations, especially considering turn-taking and addressee selection - a property partly coded in ICSI-MRDA, and a proposal for full addressee coding [199]. They discuss how SWBD-DAMSL simplified DAMSL based on frequency information but both were created for 2-party conversations (thus they might not handle MPC phenomena well), ICSI-MRDA extended SWBD-DAMSL for MPC, allowing for the combination of multiple tags (especially towards turn-taking tasks called floor-grabbing and floor-holding), and MALTUS further reduced the tagset by grouping them into classes and specifying constraints based on empirical observations. This progression aimed to capture the nuances of dialogue acts in multi-party meetings while balancing complexity and practicality. Bunt [200] (who contributed towards ISO 24617-2 for dialogue annotation, [201]) also apply the DIT++ taxonomy [202] to the AMI corpus [203, 58]. The DIT++ taxonomy offers a more fine-grained and comprehensive approach to dialogue act annotation for both two-party and multi-party conversations. It captures a broader range of dialogue intentions and pragmatic functions, considering linguistic features and incorporating cross-linguistic applicability. DIT++ emphasizes social actions and incorporates a wider range of cues, enabling a richer understanding of dialogue interactions in diverse contexts.

Furthermore, Traum et al [204] work on annotations where participants are also in

a multi-floor setting in addition to a multi-party setting, where not all have access to the visual context that others do. The proposed scheme includes a transaction unit that clusters utterances from multiple participants and floors into units according to realization of an initiator's intent, and relations between individual utterances within the unit. They showcase applications for their scheme as training and evaluation data for creating automated multicommunicators, however this data is not released for public use. Of particular interest is their discussion on how to translate taxonomies for dialogue act annotations for a scenario involving multiple participants (and social roles) and multiple floors (sub-groups which are connected yet exist separately) in the conversation.

2.5.3.3 Dialogue Act Classification

Like most other tasks, various approaches have been utilized in dialogue act classification for multi-party data.

Kim et al [205] experiment with statistical approaches such as *Naïve Bayes* (NB), *Support Vector Machine* (SVM), and *Conditional Random Field* (CRF) and propose features such as 1) bag of words - *TF-IDF* and n-grams; 2) Structural Information - distance from first utterance to target utterance, term-count, which user is the speaker and whether they are the host of the conversation depending on its context (ex in library chat settings, the librarian would be the host); 3) Keyword information and 4) Interaction among Utterances. They apply these methods towards classifying dialogue acts in the NPS chat corpus [78]. They find the use of contextual and keyword features to be useful, and suggest that entanglement amongst utterances from different participants stemming from the multi-party setting cause lower performance using structural and dialogue interaction features. Tavafi et al [206] focus on multiple corpora, including ICSI [59] for multi-party settings which consists of synchronous conversation, and compare the performance of *SVM-HMM* (which combines SVMs with HMM) and *CRF* methods towards modeling dialogue acts. While they find that

the proposed SVM-HMM algorithm with domain-independent feature set can achieve high results on synchronous conversations, they also argue that this could be a result of the reduction in complexity as a result of the sequential nature. Amanova et al [207] work on obtaining new annotated dialogue data by using available resources and supervised machine learning algorithms such as *Logistic Regression*, *AdaBoost*, and *SVMs*, introducing methods for automatic annotations and reducing costs associated with annotations for MPC using the two transcribed MPC corpora - AMI [59, 60], and SWBD-DAMSL [197]. They experiment with hybrid and fusion approaches, finding that a meta-classification strategy (which utilizes a meta-classifier that can re-classify classes by taking into account multiple datasets and models) performs better than a majority voting strategy. Irsoy et al [208] utilize neural methods using the Settlers of Catan corpus [83]. They propose using a *directed-acyclic-graph LSTM* (DAG-LSTM), where each dialogue thread is modeled as a directed (past utterance to current utterance), acyclic graph. This helps exploit the turn-taking structure naturally present in a multi-party conversation, and encode this relation in model structure. They additionally find that the results demonstrate that information about the prior utterance made by a speaker is very useful in DA classification.

In this section, we have presented tools could be useful towards transcription, annotation, and visualization of multi-party conversations, including those with multimodal data; annotation taxonomies; and classification methods which have been applied towards multi-party conversation understanding. While there are methods which have been shown to be effective, recently released pre-trained methods could be employed to improve both the annotations and classification methods.

2.5.4 Anaphora Resolution

Anaphora resolution allows more meaning-making with regards to the content of an utterance in dialogue, relating to the challenge of understanding what and how to respond. In multi-party settings, this becomes an even more interesting problem, for

example, pronouns like "you" could refer to more than one participant in the same dialogue. We discuss methods that have been used to resolve references to "you" and "it" in multi-party conversations.

Resolving "you" references. Studies involving pronoun use in multi-party conversation use were conducted as far as back as Lerner [209], who present that in multi-party conversational settings, the pronoun "you" does not automatically resolve "who" is being referred to, a challenge we discuss in detail in Section 2.2. This is an important challenge, since the pronoun resolution can help determine who the addressee is. They also propose including multimodal features such as eye-gaze to resolve the pronoun references better. Gupta et al [210] focus on resolving the referential and non-referential "you" usage in the AMI corpus [58].

They tackle the multi-party nature of the classification by modeling the problem as a 4-way classification task for each utterance and using a sequence classifier based on *Conditional Random Fields* (CRFs) [211]. It is notable that they use only text to train and test models. Frampton et al [212] focus on a similar task, also using the AMI corpus [58], similar to addressee identification work by Jovanovic and Akker [167], using a *Bayesian Network classifier* trained on multimodal features. They focus on using fewer manually annotated features, and find that eye-gaze can be highly predictive for the referential usage.

Resolving "it", "this, "that" references. Müller [213] introduce resolution for the usage of the nonreferential "it" in the ICSI corpus [59]. They utilize a machine learning system, called *Repeated Incremental Pruning to Produce Error Reduction* (RIPPER) [214] which learns to form rules for identifying the nonreferential based on features such as information about the parts of speech, and structural properties of the text. They find that including interruption points helps the classifier comes up with better rules, and thus identify the nonreferential "it" better. They further extend their study to focus on resolving references to "this" and "that" as well [215],

with future work focus on the evaluation on using the reference resolution for dialogue summarization.

Most extant research in this area, specifically in multi-party conversations, has also mainly focused on formal spoken corpora. There is thus a research gap in anaphora resolution within informal multi-party conversations, a challenging task owing to the informal nature of the data. Additionally, most approaches for anaphora resolution focus on utilizing statistical methods. Current state-of-the-art in the coreference resolution task [216] has shown that using contextual representations outperform previous approaches, and this could be an interesting direction to pursue for anaphora resolution in multi-party conversations as well.

2.5.5 Entrainment

Entrainment refers to how language used by the participants of a group changes over time to become more similar. While it is more a result of conversation than a part of the dialogue itself, multi-party settings specifically allow entrainment to be studied towards the goal of understanding negotiations and persuasion. Work in this area for multi-party conversations has been mainly focused on the TEAMS corpus [55] owing to its entrainment focused nature, except for two [217, 218].

Branigan et al [217] study the differences in how language use differs in MPC, noting that work in the area is lacking. They conduct 3 experiments varying speaker roles on a turn-based level (instead of a conversation level), and find that syntactic alignment does occur in MPC, and that it is not restricted to speaker-addressee dyads within them. However, they also find that syntactic alignment is sensitive to variations in speaker role with respect to the source utterance, but not the current addressee's role with respect to the source utterance. They conclude that it is possible for mechanisms of syntactic alignment to differ in some respects from the mechanisms of lexical and semantic alignment, requiring some elaboration of the interactive-alignment model [163] towards MPC modeling.

Friedberg et al [218] study student group success in academic environments, arguing that higher entrainment in groups could point to a shared mental model, which in turn could point to better success with their projects. They use the pair score [219], to determine the entrainment scores by measuring the similarity in the use of high-frequency words between two speakers. The experiment showed that higher scoring teams are more likely to increase their entrainment in project words over the course of a conversational session, while lower scoring teams are more likely to diverge in their use of project words. The study highlights important challenges faced when dealing with multi-party corpora - considering the team sessions separately leads to a significant result whereas just focusing on two-party interactions does not.

Rahimi [220] utilize the TEAMS corpus [55] to study entrainment for non-dyadic measures. They utilize two conversation-level measures of entrainment: *proximity* (degree of similarity between members within a team relative to participants in other teams) and *convergence* (change in similarity of teammates over time) by averaging corresponding dyad-level measures. They find that entrainment occurs on two levels - acoustic-prosodic and lexical - and that both these scores positively correlate in both directions. They also find that to predict positive and negative team outcomes, a multimodal model with features from both levels of linguistic entrainment outperforms a unimodal model, highlighting how multimodal corpora could contribute towards advances in multi-party analyses. Yu et al [221] instead focus on features of the speaker's linguistic style using LIWC [222] by measuring the "function" words, the team's linguistic style entrainment using the methods described by the TEAMS paper [55], and team characteristics such as demographic data, to predict perceived team social outcomes such as conflict along with entrainment. They find that teams with greater gender diversity had greater minimum convergence than teams with less gender diversity, and that 4-person teams with more than one female had a higher maximum magnitude of change in team difference. In terms of team characteris-

tics, they find that the maximum magnitude of the change of the team difference was negatively associated with team conflict, and that entrainment was significantly negatively associated with task and process conflict, both when controlling for team characteristics and when not. Inclusion of several multi-party properties makes this study an interesting step towards predicting how entrainment contributes to group success. Rahimi and Litman [102] introduce learning entrainment vectors using neural methods, calculated by converting dyadic entrainment measures [223], then utilizing various graph similarity measures (called kernels) such as closeness centrality [224] and PageRank [225] to create weighted graphs, and then make the representations denser similar to methods used for building `word2vec` [226]. They propose three such methods to learn group entrainment embedding: direct estimation with graph centrality kernels, self-supervised approach using autoencoders, and weakly supervised approach using labelled data. They find that the weak-supervision approach where the denser representations are optimized towards predicting whether the entrainment graphs given as input are real or permuted, outperforms previous baselines.

Studying entrainment allows research to understand how multi-party settings affect the participation levels as well as linguistic properties of dialogues in these settings. including goal-setting for dialogue agent participation in multi-party conversations. One such dilemma is that dialogue systems are not yet trained to lexically entrain towards human participants, which means that the entrainment falls on the humans, and is an observed effect. Another dilemma lies in the presence of multiple participants in the dialogue, since a dialogue system would need to learn to entrain for one or more participant in the conversation, depending on the context under which it is participating in the conversation. Studies have shown that there might be differences based on not only conversation level, but turn level context, another important property which needs further investigation.

2.5.6 Topic Identification

Since multi-party conversations carry with them the possibility of multiple topics occurring concurrently during the same conversation, extant research has focused on modeling topics within multi-party conversations. Kim and Baldwin [227] focus on retrieving relevant keywords from live chats. They use structural information such as predicted dialogue acts, which are determined using features such as stemmed bag-of-words, highly frequent terms, and participant information, to find relevant keywords. They postulate that this introductory work will contribute towards better topic modeling in future work.

Quite a lot of methods utilize the ICSI corpus [59]. Purver et al [228] present an unsupervised topic modeling method adapted from document classification [229, 230] based on Bayesian inference for generating topics towards automatic topic segmentation and identification, finding the performance comparable with well-rated extracted topics based on human evaluation.

Georgescul et al [231] present a comparative study of two probabilistic mixture models based on Latent Dirichlet Allocation (LDA) [229] and Aspect Model for Dyadic Data (AMDD) [232], finding both models to perform comparably with previous state-of-the-art methods. Nguyen et al [193] present the Speaker Identity for Topic Segmentation (SITS) model, based on a hierarchical Bayesian nonparametric model, for discovering 1) the topics used in a conversation, 2) how these topics are shared across conversations, 3) when these topics shift, and 4) a person-specific tendency to introduce new topics. They find that while the model performs well, it particularly helps while modeling debate conversations, stating that a possible reason could be that speaker identities are more pronounced in debate.

Most methods adopted by the papers in this section utilize Bayesian methods, which is reflective of popular modeling techniques within topic modeling. The focus seems to have lingered on formal and debate corpora, and while these are naturally

expected to have a topical structure to them, informal discourse also revolves around topics and would be an interesting area to look at.

2.6 Current Advances in Multi Party Conversational AI

An overview of some of the most challenging aspects and considerations for multi-party conversation settings have been presented thus far. In this section, we focus on conversational systems which include dialogue management for multi-party conversations.

2.6.1 Modifications to existing two party based systems

Nishimura et al [233] focus on extending a two-party (1 user 1 agent) system to work for a 3-party (1 user 2 agents) system, an active-additive (type 1) system. They find that including multiple agents allows them to model differing viewpoints, making the user experience more enjoyable when conversing with them. They focus on fixed topics for conversation, which limits the domain knowledge necessary. The dialogue manager consists of five sub-components, consisting of 1) *Information collection* which consists of obtaining linguistic cues from the SPOJUS speech analyzer [234] output, 2) *Feature extraction* which calculates response timing and response type using *decision trees*, 3) **Response generator** which uses template matching and considers dialogue context using slot information since the topic of the dialogue is known, 4) *Response timing generator* using prosodic features again with a *decision tree*, based on their previous work [235], and 5) *History manager* which saves the conversation history for dialogue context management. They observe that the 3-party system performs better in regards to "familiarity with the agent", "interest in the topic", and especially, "easy to speak to" , "various opinion", "lively conversation" and "like chatting"; for general chitchat.

Hu et al [174] present a *Graph-Structured Network* (GSN) since dialogue **response generation** assume that utterances are sequentially organized, but multi-party con-

versations have the added challenge of the possibility of utterances not being so. GSN consists of three stages for dialogue generation: 1) *Word-level encoder* using a bidirectional *recurrent neural network* (RNN) with *Long Short Term Memory* (LSTM) units to encode each word, 2) *Utterance-level Graph-Structured Encoder* (UG-E) using a *Hierarchical Encode-Decoder* (HRED) which is a hierarchical sequence-based word and utterance-level *RNN*, then both these stages are explained as matrix manipulations, and lastly 3) *Decoder* to generate the response using a *Gated Recurrent Unit* (GRU) network. They evaluate their approach on the Ubuntu dialogue corpus [79], as an active-simulative (type 2) system, , finding through ablation studies that their approach performed best, and that GSN works well for both two-party and multi-party dialogue.

Martinez and Kennedy [236] present a dialogue system with concurrent conversation tracking and memory, with the goal being the design and implementation for missing pieces required to leverage two-party dialogue systems in order to provide a more natural multiparty conversational experience. Towards this end, they develop a new sub-module to be included into the dialogue manager, called a concurrent conversation manager. It keeps track of topics and participants in each of the conversations and sub-conversations as they arise in the dialogue, as well as identify opportunities for the agent to connect with particular members of the group. They focus on **addressee recognition** and include the task of one-to-one vs one-to-many communication within multi-party conversation. They approach this problem using dialogue graphs, and possible paths forward are determined by probabilistic score. This active-additive (type 1) system is tested with real life participants to maintain a lively conversation. They find that the model performs well, showing that there is the possibility of extending two-party systems to multi-party settings. However, they also find that thread management with more than seven participants can cause problems for the system due to the comparatively lower pace at which the system can

follow the conversation compared to humans, pointing to future work.

2.6.2 Component based multi party systems

The CALO Meeting Assistant (CALO-MA) [237, 238] was probably one of the first complete conversational systems that was built for automatic transcriptions and annotations as a passive (type 3) type system, focusing on meeting data. It offers a wide range of capabilities for real-time and offline speech transcription, consisting of dialog act segmentation and tagging, question-answer pair identification, action item recognition, decision extraction, and summarization, for meetings conducted online via the Internet. They utilize the SRI-ICSI NIST meeting recognizer [239] which performs a total of seven recognition passes, performing online speech detection, causal feature normalization and acoustic adaptation, as well as sub-real-time trigram decoding. Next, CALO-MA performs dialog act segmentation with DAMSL [196] and MRDA [60] meeting annotations, using hybrid models combining *hidden event language models* (HELMs) with a discriminative classifier, namely *Boosting*. An intermediate step the system takes is **speaker identification** and **addressee recognition**, which is done by analyzing linguistic cues such as pronouns [210], and using these features for a *Conditional Random Field* (CRF) to model discourse context, explicitly modeling forward and backward dependencies in the dialog. Next comes topic identification and segmentation, utilizing *Latent Dirichlet Allocation* (LDA) for multi-party interactions [228]. Then action items and decisions are extracted by taking a structural approach to detection: utterances are classified according to their role in the commitment process (e.g. task definition, agreement, acceptance of responsibility), and then action item [240] or decision discussions [241] detected from patterns of these roles. Finally, they perform meeting summarization following Riedhammer et al [242]. Thus, they present a holistic system of analyzing multi-party meetings.

de Bayser et al [243] present a comprehensive study, which details the challenges for multi-party dialogue systems, categorizes the state-of-the-art for each challenge

in 2017, and present a hybrid conceptual architecture along with insights into lessons learned when it is deployed for the finance domain. They present a conceptual architecture for multi-party aware chatbots. Their text-related workflow consists of 1) an optional *topic classifier* which is domain-dependent, 2) *dependency parsing* (using *Speech-Act Based Intelligent Agents*, or SABIA, realized via Akka [244]) for identifying semantic dependencies, 3) *frame parsing* which helps identify slot-filling frame for recognized intents (to be implemented independently), 4) *intent classifier* using *1-nearest-neighbour* and *SVMs*, 5) *speech act classifier*, and 6) *action classifier* using SABIA as well. They realize this system for the finance domain, called Cognitive Investment Advisor (CognIA), as an active-additive (type 1) system, and the authors note that further development in underway.

Multi-role Interposition Dialogue System (MIDS) [245] focuses on **response generation** based on dialogue context and next speaker prediction. MIDS employs multiple role-defined encoders to understand each speaker and an independent sequence model to predict the next speaker which as a scheduler to integrate encoders with weights, consisting of 3 types of *RNNs* to encode input data for different tasks - 1) a set of role-number encoders (*LSTMs*) that encode each corresponding roles' speech; 2) an extra contextual *RNN* (*bi-LSTM*) that encodes all speakers' speech; 3) a speaker *LSTM* that encodes speaking order. Then, an attention-enhanced decoder generates responses based on dialogue context, speaker prediction and integrated encoders. Moreover, with the help of the unique speaker prediction, MIDS is able to generate diverse responses and join conversation actively when appropriate, and interact with users without cue during real-life online conversations. It is tested on the Friends dataset, as well as in real life on the WeChat [246] platform as an active-simulative (type 2) system.

ConvLab 2 [247], built upon its predecessor [248], inherits its framework and extends it by merging several recently proposed state-of-the-art approaches. Its main

components consist of: 1) *Dialogue Agent* allowing multiple configurations such as pipeline, end-to-end systems, and even between other layers (once they are instantiated); 2) *Models* which consists of sub-components such as 2.1) NLU (supports Semantic Tuple Classifier, or STC [249], Multi-Intent Language Understanding, or MILU [248], and BERTNLU proposed by the authors based on *BERT* [250] and adding two *Multilayer Perception*, or MLP layers), 2.2) Dialogue State Tracking (a rule-based tracker for belief state updates), 2.3) Word-level Dialogue State Tracking (ConvLab-2 integrates three models here - Multi-Domain Belief Tracking, or MDBT [251], Slot-Utterance Matching for Universal and Scalable Belief Tracking, or SUMBT [252], and Transferable Dialogue State Generator, or TRADE [253], 2.4) Dialogue Policy (rule-based policy, a simple neural policy that learns directly from the corpus using imitation learning, and reinforcement learning policies including REINFORCE [254], Proximal Policy Optimization algorithms, or PPO [255], and Guided Dialog Policy Learning, or GDPL [256], 2.5) Natural Language Generation (NLG) (template-based method as well as Semantically Conditioned-LSTM, or SC-LSTM [257], 2.6) Word-level policy (itegrates three models - MDRG [258], Hierarchical Disentangled Self-Attention, or HDSA [259], and Latent Action Reinforcement Learning, or LaRL [260], 2.7) User Policy (an agenda-based model [261] and neural network based models including Hierarchical User Simulator, or HUS and its variational variants [262] and lastly 2.8) and End-to-End Model (ConvLab-2 extends Sequicity [263]) to multi-domain scenarios, and integrates Domain Aware Multi-Decoder, or DAMD [264] and ROLLOUTS RL policy [265]; 3) support for loading multiple *Datasets*; 4) *Analysis* tool for evaluation; and 5) *Interactive Tool* which provides a graphical interface to interact with the dialogue system. While the datasets they include do not have multi-party focused ones, the paper does mention that ConvLab-2 can manage multi-party dialogue as well, possibly as both an active-additive (type 1) and active-simulative (type 2) system. Interestingly, a new version has been proposed [266], but there is no

mention of support for multi-party conversations.

The *Plato Dialogue System* [267] consisting of 4 modules: 1) *Dialogue* component for **dialogue management**, 2) *Domain* for intent slot-filling in task-oriented dialogue, and 3) *Controller* for orchestrating the conversations between agents towards **turn-taking** - these could all be rule-based modules or trained modules based on various machine learning frameworks, which means the system allows swapping out components. The *Agent* module embodies a pre-defined role (such as travel agent) which is realized by components with NLU, dialogue management, dialogue state tracking and NLG. This role can have different definitions, supporting both behaviors - active-additive (type 1) and active-simulative (type 2).

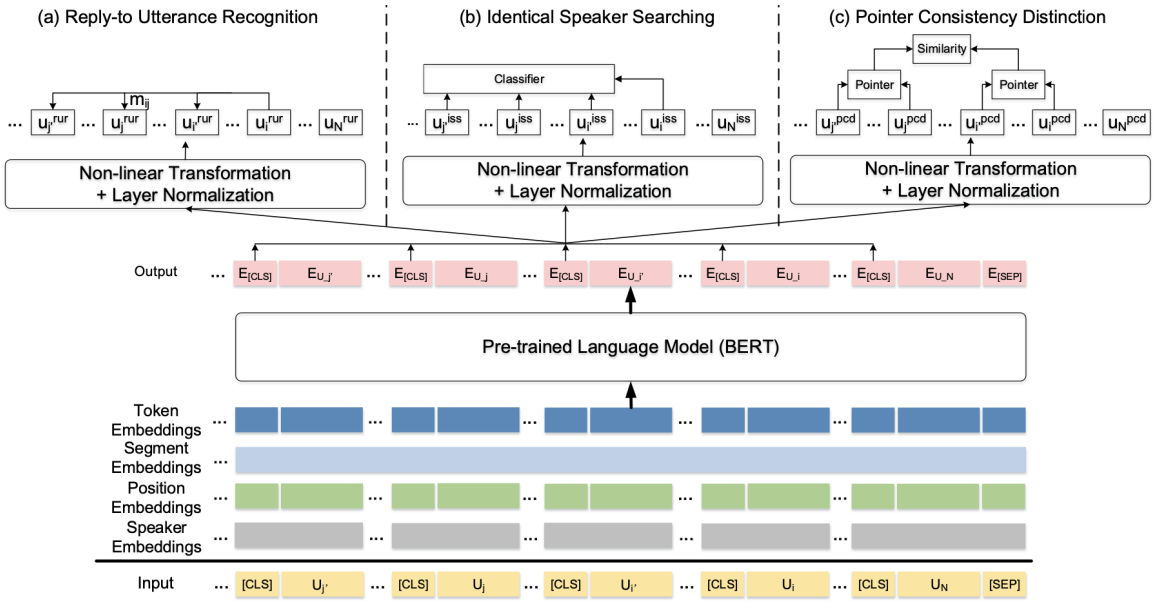


Figure 2.7: Input representations and model architectures of the three self-supervised tasks for interlocutor structure modeling, including (a) reply-to utterance recognition, (b) identical speaker searching and (c) pointer consistency distinction [12].

MPC-BERT [12] present a pre-trained model based on BERT [250] for multi-party chat understanding that considers learning who says what to whom in a unified model (Figure 2.7) with self-supervised tasks called 1) *Reply-to Utterance Recognition* to learn which preceding utterance the current utterance replies to, 2) *Identical Speaker Searching* to search for the utterances sharing the identical speaker so as to identify the

speaker of an utterance, 3) *Pointer Consistency Distinction* for identifying a pair of utterances representing the "reply-to" relationship (defined as a speaker-to-addressee pointer), 4) *Masked Shared Utterance Restoration* to identify shared utterance that are semantically relevant to more utterances in the context than non-shared ones, and 5) *Shared Node Detection* which utilizes the conversation structure to strengthen the capability of models on measuring the semantic relevance of two sub-conversations; modeling speaker role identification and response selection tasks (building on their previous work [178]) within one system. Experimental results on three downstream tasks show that MPC-BERT outperforms previous methods by large margins and achieves new state-of-the-art performance on Ubuntu IRC data [79] as an active-simulative (type 2) system.

Approaches for modeling multi-party conversations have thus utilized statistical and neural approaches, and consist of components similar to two-party dialogue systems, with further provisions to handle multi-party specific challenges. It is interesting that approaches take both routes - modifying two-party systems and building multi-party systems specifically. However, there is a need to focus further especially for open-domain dialogue generation for multi-party dialogue, and for an evaluation benchmark that could take multi-party specific challenges into account which would allow better comparisons across existing and future research - evaluation considerations which we discuss in Section 2.8.

2.7 Applications for Multi Party Conversation Systems

We discuss the various applications that multi-party conversation research has found in real life, and how it has enriched the interactions between humans. The applications of such systems range over multiple use-cases, showing that multi-party systems could greatly improve the quality of group conversations in society.

Game-playing. Morgan et al [268] introduce an agent for providing mentoring to facilitate progress and learning in a serious game setting as an active-simulative (type

2) system. They categorize player and mentor contributions into eight different speech acts and analyze the sequence of dialogue moves using State Transition Networks. They find that even when the transition frequencies between the task- and discussion-oriented stages were overlapping, they were able to observe that mentor requests and player questions reflected the goal-driven activities of the task-oriented stages, whereas the discussion-oriented stages showed greater emphasis on player statements and expressive evaluations as they reflected on previous game actions. The feedback was more likely to be provided by the mentor in the task-oriented stage (indicating scaffolding), but in the discussion format, other players were increasingly likely to respond (suggesting collaboration). They plan future studies to replicate results for an automated mentor.

Social robot interaction. Kenny et al [269] present a demo of interaction with virtual humans, including features such as multi-party negotiation. The general architecture of the presented system consists of 7 major components, of which 3 relate to the multi-party text communication: 1) *Speech Recognition*, 2) *Natural Language Understanding*, 3) *Intelligent Agent* which reasons about plans and generates actions using the Soar Cognitive architecture [270]. They also include components towards dynamically generated non-verbal behavior in the virtual environment, but we focus on behavior within the scope of this survey. While the discussion in the paper is limited to the system itself (which can handle both active-additive type 1 and active-simulative type 2 interactions), the demo showcases interesting applications in virtual interactions which have been garnering more interest in the software industry lately [271].

Hartholt et al [272] introduce an interactive virtual environment based on a Wild West setting, called Gunslinger, as an active-simulative (type 2) system. The project combines virtual humans technology with Hollywood storytelling and set building into an engaging, mixed-reality, story-driven experience, where a single participant can

interact verbally and non-verbally with multiple virtual characters that are embedded in a real, physical saloon, allowing multimodal data collection in multi-party settings. Along similar lines, Foster et al [273] showcase a humanoid robot bartender that is capable of dealing with multiple customers in a dynamic, multi-party social setting as an active-simulative (type 2) system. Their system consists of visual processing, linguistic interaction, state management, and high-level planning and monitoring. They find that the bartender is able to handle a range of social situations, showing yet another interesting application for a social multi-party system. An application which aims to study the effect of group size on human-robot interaction is presented by Leite et al [274], comparing individual vs group conversations, a passive (type 3) analysis. They measure disengagement using SVM-based models to training with data from participants interacting alone with two social robots vs participants interacting with the robots in small groups, and a third model combining data from the two datasets. They find that a model trained with group data generalizes better to individual participants than the other way around, again showcasing how useful multi-party systems could be in casual communications. Candello et al [275] even recreate artwork experience based on "Café com os Santiagos" by three Brazilian artists as active-simulative (type 2) system: Claudio Pinhanez, Heloisa Candello and Paulo Costa with multi-party systems - describing a unique setup allowing interactions with chatbots in space recreating a 19^{th} century coffee table, showcasing creative applications.

Application in the health domain. Snaith et al [276] present a dialogue game in which two or more health coaches collaborate with a patient to help with goal-setting for behavior change using both active-additive (type 1) and active-simulative (type 2) settings. They focus on guidelines to approach discussions, persuasion, and goal-setting in multi-party conversation scenarios. They even provide an open-sourced project on Github called AgentsUnited [277] for virtual agent interaction surrounding health [278], noting the popularity of their proposed methods in current ongoing

research projects. Fioramonte and Vásquez [39] focus on healthcare settings where workers also meet additional participants aside from the patient, such as family members, especially in pediatric and geriatric contexts. They closely examine the linguistic and discursive strategies family members employed when actively participating in helping provide more details about patients' conditions or health history. This setting showcases how important it could be to conduct research involving multiple participants, allowing for important perspectives to be brought into health conversations as active-additive (type 1) systems. Das et al [279] apply multi-party dialogue generation towards promoting active living and healthy dieting for type-2 diabetes subjects, through multiple embodied conversational agents. They create a set of virtual coaches, specialized in different areas such as exercise training and nutrition, who interact with a patient to provide support and help them adopt a better lifestyle. The authors propose further studies to evaluate the effects and user experience. Thus they showcase both active-additive (type 1) and active-simulative (type 2) system settings.

Driving assistance. Karatas et al [280, 281] introduce a multi-party conversational social interface NAMIDA through a pilot study, consisting of three robots that can converse with each other about environment throughout the drive. Through this model, the directed utterances towards the driver diminishes by utilizing turn-taking process between the agents, and the mental workload of the driver can be reduced as compared to direct communication with the driver. They find that the trend analysis demonstrated that their proposed multi-party conversation-based active-additive (type 1) system is promising in reducing the attention behavior on the system over use. Furthermore, they propose future work to consider the instant condition of the driver, behavior and workload-wise.

Furhat. Probably one of the most well-known applications for multi-party conversations has been Furhat [40, 41]. First introduced as an active-simulative (type 2) system, Furhat has since come a long way, with over 180 publications related to it.

It is a multimodal system, using features such as speech recognition, eye-gaze, and head movement [282] for addressee detection and response selection [283], as well as cues for learning turn-taking behavior [284, 285].

Multi-party conversation agents have thus found use in multiple domains where they are able to facilitate interactions with group conversations. While most applications are in active systems (mostly simulative), research in more passive areas is also gaining traction with multiple workshops being organized towards dialogue summarization [286] and meeting transcription [287], extending to meeting action and decision summarization [288], although these application are beyond the scope of this survey. Overall, MPC systems have even been shown to be more effective than dyadic interactions of a user with a single conversational agent, motivating the need for further research in this area.

2.8 Evaluation Considerations

Howcroft et al [289] present the challenges the NLP community, and in particular the multitude of research on Dialogue Systems, have faced in the area of human evaluations - showcasing the confusion created by the absence of a benchmark evaluation standard that covers both automatic and human evaluation methods. In recent times, there have been efforts fill this gap and introduce standardized benchmark metrics to be able to centralize evaluations and make comparability between proposed dialogue systems easier, like the GEM benchmark [290]. However, as discussed in Section 2.2, multi-party conversations introduce their own set of challenges, which need to be accounted for while evaluating multi-party systems. This section brings together some proposed discussion for the evaluation of multi-party conversations, since most currently popular metrics do not account for challenges specific to multi-party interactions. Similar to the confusion in dialogue systems, multi-party evaluation proposals have been quite disconnected. While there are subjective discussions focusing on specific task-based evaluations, we only found one benchmark which formalizes response

selection with multi-party conversation tasks [291]. We first discuss the subjective discussion with evaluation considerations proposed by both papers in Figure 2.8, color-coded in a similar fashion to the challenges from Figure 2.2, and then present the benchmark.

Dignum and Vreeswijk [13] work towards creating a testbed, mentioning multiple challenges that become a part of multi-party conversations which need to be accounted for while evaluating systems.

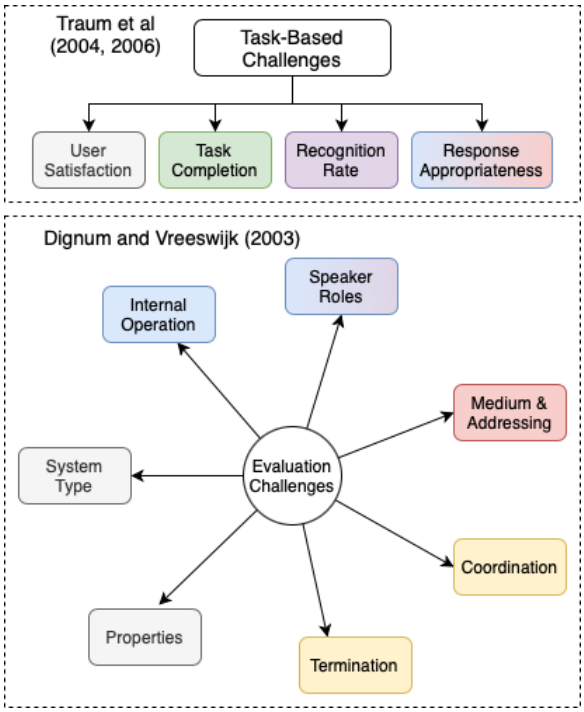


Figure 2.8: Evaluation considerations proposed in existing research [13, 14, 15], color coded similar to Figure 2.2 to highlight the connections with previously presented challenges for multi-party conversation systems.

They discuss general multi-party dialogue evaluation considerations for:

1. **System Type** - open vs closed - can participants come and go while the conversation progresses?

2. **Speaker Roles** - speaker roles by type with examples:

   (a) *Linguistic* - speaker, addressee, auditor, overhearer and eavesdropper

   (b) *Dialectic* - neutral party, interested party, interviewer, advocate, respondent, examinator, challenged party, mediator, or arbitrator

   (c) *Social* - chairperson

   (d) *Interests* - negotiating for profit, terminating early/late

3. **Medium & Addressing** - one -one vs one-many vs broadcast

4. **Coordination** - concerning aspects:

   (a) React sync/async?

   (b) Turn taking?

   (c) React to all messages or only where addressee is present?

   (d) How to they react?

5. **Termination** - when should the dialogue end? Some concerns:

   (a) In negotiation or persuasion - does dialogue end when everyone is satisfied?

   (b) Who determines termination?

6. **Properties** - these are very contextual, but some examples:

   (a) Can we guarantee that an inquiry dialogue protocol will deliver the answer if the union of all the knowledge of the parties in the dialogue would be enough to derive this answer?

   (b) A protocol only reveals information to participants that they need to know in order to respond or whether information is released that parties would rather not divulge if not needed.

   (c) Guaranteed termination?

7. **Internal operation** - Any extra properties, such as some responses being much more suitable than others given the group setting?

code	description
PF	filled pause
RR	request for repair
AP	appropriate response
INI	appropriate new initiative
CON	appropriate continuation
NAP	inappropriate response,initiative or continuation

Figure 2.9: System Agent Appropriateness Codes [14, 15] for labeling each utterance for response appropriateness.

For task-based multi-party systems, Traum et al [14, 15] present the following measures to keep in mind:

1. **User satisfaction** - were the users satisfied with the outcome of the group negotiations? Have all participants been able to express their needs and make meaningful comparisons based on the discussions?

2. **Intended task completion** - did the dialogues aid in completing the intended task? Were all participants able to express their intentions?

3. **Recognition rate** - were the addresses recognized well? Similarly, were the dialogue acts, domain concepts, and intent classification?

4. **Response appropriateness** - were the interactions natural in the context of the conversation? For example, rejections and negotiations might be appropriate in a negotiation or persuasion setting. They propose an annotation scheme for this evaluation in particular, presented in Figure 2.9.

Traum et al [292] also present a framework modeled on their SASO-EN corpus [139], which focuses on virtual human interactions with a few text-based components. They describe a real-time system which can similarly incrementally listen, interpret, understand, and react to what someone is saying, based the Listener Feedback Model [293]. The components related to text that their system consist of are: 1) *speech recognizer* which provides word-by-word transcripts with confidence scores using PocketSphinx [294], 2) *Natural Language Understanding* (NLU) and *meta-NLU* component which

produces semantic representations and predictions of final meaning and can handle partial recognition [295, 296, 297, 298], 3) *dialogue manager* [299] and *domain reasoning* [300, 301] which can update state and calculate communicative intentions. Perhaps the most interesting component is one which focuses on **participant roles** - it finds the conversational role for the participants (such as active-participants, overhearer etc), and the utterance role (such as speaker, addressee etc) using algorithms based on previous work [302, 303]. However, evaluation is reserved for future work.

The DSTC8 Track 2 (NOESIS II) [304] sets up a benchmark for response selection in multi-party conversation [291]. The approach focuses on building systems based on the Ubuntu IRC dataset [305]. The task defines the input as (context, candidates for response+speaker pair) and the output as (selected response+speaker pair). Since this is effectively a classification task, evaluation for the task uses recall and mean reciprocal rank along with precision, accuracy and F-scores for sub-tasks.

This task brings to fore the fact that most of the core tasks within multi-party conversations could be formalized as classification tasks for automatic evaluation. However, as discussed above [13, 14, 15], there is also the need to consider the properties introduced by multi-party participation, and there is a need to work towards a more holistic evaluation benchmark.

2.9 Discussion

This survey motivates the need for future work in multi-party conversation modeling by looking at the various challenges posed when multiple participants take part in the conversation. We detail the corpora which contain multi-party conversations, and further discuss data collection methods and annotation tools as a reference for future data collection in this area. We then present the various qualitative approaches taken towards studying tasks within multi-party conversation modeling. We tie together various research directions which have focused on tasks within multi-party conversation modeling, as well as end-to-end systems which either utilize the components from

task-specific models or attempt end-to-end modeling. Lastly, we motivate the need for further research in multi-party conversation modeling by looking at the various applications that multi-party conversations have already contributed to. Furthermore, we discuss evaluation considerations that need to be made to take challenges pertaining to multi-party conversations into account.

Multi-party corpora have thus been collected in multiple ways, and have differing properties which could be useful for different kinds of modeling. Unscripted spoken corpora are increasingly collected in the formal domain, and there is a need to ethically collect informal, spontaneous conversations. Scripted corpora are becoming increasingly available especially as movie and TV transcriptions. For written corpora, social media platforms have become significant sources of data, however there is a need to be able to collect these with the required properties and format them towards modeling tasks. With all these categories, there is still a need towards collecting datasets which contain properties such as interlocutor persona, overhearers, conversational environmental conditions, etc.

Approaches for modeling multi-party conversations have thus utilized statistical and neural approaches. It is interesting to note that approaches take both routes - modifying two-party systems and building multi-party systems specifically. However, there is a need to focus further especially on open-domain dialogue generation for multi-party dialogue as modes of conversation evolve with time (as evidenced in light of the COVID-19 pandemic when most communication had to evolve towards utilizing the internet and the main mode of communication for personal as well as work communication). From the system type point of view, most research has focused on active type systems, specifically more so on active-simulative type systems. This is probably in direct correlation with the multiple challenges that multi-party conversations introduce, and how they compound towards making AI participation difficult - replacing a present participant seems an easier goal to achieve, compared to creating

a new participant, since the data for training a model already exists. Applications for all system types have been shown to be useful (Section 2.7).

Concurrently, there is a need to contribute and study more recent corpora with MPC, especially written corpora, which can provide more insights into evolving trends in language and usage in group-based communication. We discuss this need further in Chapter 3, and contribute a mock social media platform towards tools for data collection.

Given that the presence of multiple participants have been shown to affect response generation, there is a need to study the effect of user properties such as persona on the same. We study this and present our findings in Chapter 4.

Additionally, there is a need for an evaluation benchmark that could take multi-party specific challenges into account which would allow better comparisons across existing and future research. We discuss the evaluation considerations towards MPC in Chapter 5 for a better understanding of progress in this area, and discuss shortcomings which need to be overcome.

CHAPTER 3: Corpora for Multi Party Conversation Modeling

3.1 Introduction

Existing corpora for multi-party conversations span the range of informal and formal corpora, with spoken corpora being more prevalent than written corpora (Section 2.3, Table 2.3). Our survey [182] finds that this property is similar when compared to previous surveys (which cover mainly two-party corpora) [306]. However, the nature of multi-party conversations is more prevalent on the personalized web, and thus social media sites and forums (such as Twitter [138], Reddit [131], and Ubuntu IRC [129]) can to be great data sources towards building MPC corpora, as evidenced in Figure 2.3. Additionally, we aim to collect more features which could help further research in multi-party conversation modeling, such as persona attributes for the users.

While there is no lack of freely available written text with the advent of the personalized web, there are limited corpora that have been extracted from these sources. Additionally, the limitations of available data are exacerbated by the lack of relevant information, which makes it difficult to utilize them for MPC modeling. Most of these existing written corpora come from specific research efforts which aim to produce multi-party conversation corpora, since building corpora often requires processing to ensure presentation in a suitable format for modeling tasks (Speaker Identification, Addressee Recognition, Response Selection/Generation). The need for research stems not only from the need to obtain consent to record data and maintain the privacy of the participants, but also from the ever-changing nature of the platforms which host the data and the requirements to comply with their Terms of Service.

One way to satisfy this need is to introduce mock platforms which allow more control over permissions as well as data collection. Several efforts have been made

in this area, towards both synchronous [82] and asynchronous [307] tools for written corpora.

Thus we introduce Community Connect - a mock social media platform with a user interface similar to Twitter, which has been shown to be the most popular platform for academic research [308]. In addition to providing basic data collection features (posts, replies, reposts, likes, etc), the platform also provides group-based divisions. This means that each user belongs to a specified group, and allows for "bridge" users who can help posts to travel across groups. This important property lends to two main features which are important for MPC - 1) tracking the "overhearers" since the audience for each conversation is known to everyone who posts something on the platform (via the Connections pane) and 2) understanding how users with similar ideologies can respond differently when the audience has different properties (e.g. how being in a homogeneous liberal group vs in a heterogeneous liberal + conservative group affects what people post).

We also describe our data collection efforts from user studies conducted via Community Connect. We make the code utilized for converting this data into a formatted dataset (which contains properties similar to the Ubuntu Chat corpus [79] as described by Hu et al [174]) publicly available. We describe how the format relates to the information which we collected via Community Connect, and statistics of our final dataset. Our final dataset is a formatted version of the data collected via Community Connect. Along with this, we also contribute persona attributes which were 1) collected separately via user surveys and 2) generated via zero-shot prompting based on user behavior (Section 3.5.2).

3.2 Motivation

There have been several projects developed to conduct computational social science experiments that emulate existing social media platforms. We surveyed 11 projects publicly available on Github and in published articles, summarized in Table 3.1.

Table 3.1: Comparison of repositories and projects of existing mock social media platform from Github as of Mar 2021.

Platform Name	**Similarity with mainstream platforms**	**Last Updated**
Social Lab [309]	Facebook	2014
YourNet [310]	Twitter	2017
Fireblogger [311]	Twitter	2017
Teal [312]	N/A	2017
AntiSocial [313]	Facebook	2018
FaceBird [314]	Twitter	2019
Truman [307]	Twitter	2019
iSocial [315]	Instagram	2020
Friend.ly [316]	Combination	2020
Bloom [317]	App	2020
Spruce [318]	Facebook/Google+	2020
Community Connect [153]	Twitter	2021

Most existing work had complete projects with documentation in their Github repositories. Several platforms had a look-and-feel similar to Twitter, while there were a few with similarities to Facebook, Instagram, or a combination of various platform interfaces. The repositories were also up-to-date; while several had not been updated within the last year. However, one key functionality - the ability to put users into specific groups that allow limiting the kind of interactions that are possible within and outside the groups - was missing from all existing platforms.

This key drawback motivated the need to develop our own social media platform. To the best of our knowledge, Community Connect is the only platform to design experiments of controlled information flow across groups through bridge nodes. While we emulate the look and feel of Twitter, we did not reverse engineer any of its functionality, and instead provide our own representation of structured data that can be collected from Community Connect.

We built Community Connect to facilitate data collection within a controlled environment, and allow more control over defining groups that will interact with each other. Thus, not only is the platform able to collect multi-party corpora, but it is also able to study how content from travels from one disconnected group to another through *bridge* users, i.e. users that are members of both groups [319]. What moti-

vates bridge users to share content and how does it affect the members of the groups they belong to, as measured by their engagement (e.g. likes, shares)? Our work thus makes several salient contributions:

1. Community Connect is an *open-source platform*, modeled after Twitter, the most popular platform for academic research [308], to make it easy to use for participants.

2. The code is designed to be *customizable and scalable*, enabling the research community to adapt it easily.

3. *Researchers can limit user interactions* based on *group* access, such that members of one group may not access posts from groups to which they do not belong.

3.3 Platform Overview

We discuss details about the platform which serve as additional documentation towards understanding how to utilize and modify it. We keep MPC corpora collection guidelines [182] in mind while we create the platform.

3.3.1 Technical Implementation

We utilize the MEAN stack - MongoDB [320] for storing data, ExpressJS [321] and NodeJS [322] for server-side script writing and building APIs, and AngularJS [323] for building the user interface which participants interact with. The MEAN stack [324] is a popular framework to build applications on, since the entire web application utilizes the JavaScript language.

The main service is hosted on an Amazon EC2 instance. We use a load balancer in front of the main service to make the service scalable and consistent. The data is organized by (a) *user* collection, which contains user information including which group they are a part of; (b) *feeds* collection, which contains all the posts made by each user along with metadata such as timestamp and the engagement fields (number

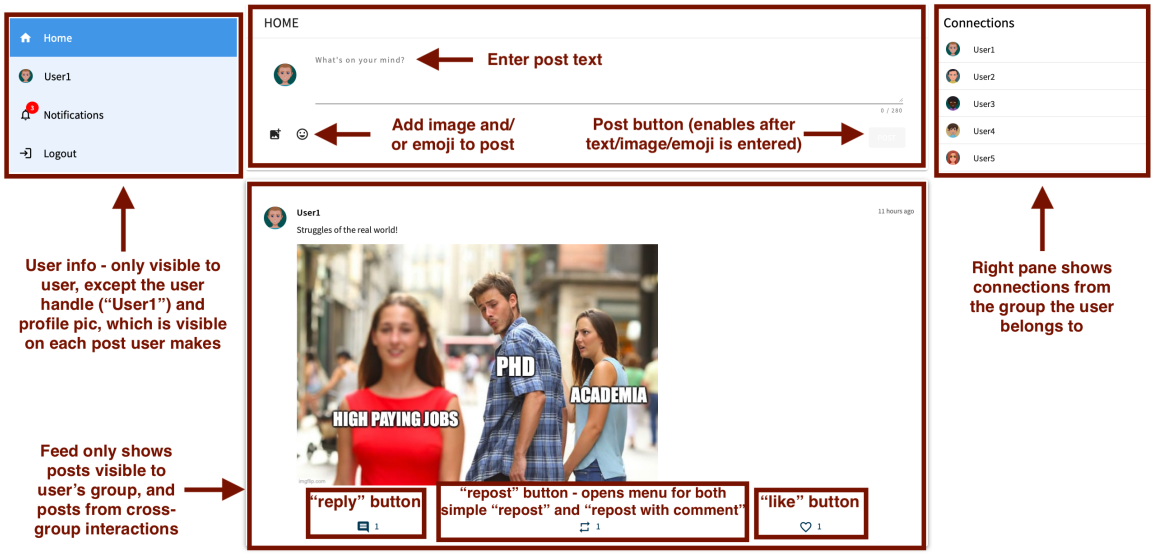


Figure 3.1: Main interface of the Community Connect social networking platform and key features that are supported, including posting, reposting, quoting, liking, and replying.

of likes and replies a post has received); and (c) the *conversation visibility* collection, which records the initial visibility as to which post is visible to which group. It also maintains records of how the visibility of each post changes when it travels across group via bridge users. An additional *notification* collection records important actions such as replies to provide functionality for a seamless user experience for referring to posts.

3.3.2 Interface Design

Figure 3.1 shows the main interface of Community Connect, which has been designed to emulate the affordances of the popular social media platform, Twitter. Participants have the ability to post images and emojis along with their text content. They can also interact with content by *liking* the post, sharing the post as-is (*reposting*), sharing the post while adding their own thoughts on it (*quoting*), or *replying* to the post. The user profile information is found on the left-most pane, including their user name (*purple_rain*) which opens up a page to change their password should they

wish to do so. This pane also includes a Notifications section, where notifications for engagement on the user's posts can be found and clicked on for easier post navigation. Messages can be liked, shared, quoted, and replied to, as seen in Fig 3.1. On the right-most pane, the user is able to see all the other users they are connected to, which could belong to one or more groups. This allows a user to navigate to the posts made by a specific user easier. Since each user is only connected to defined groups, they can see who the people they can respond to regularly are, using this pane.

3.4 Data properties captured via Community Connect

Community Connect stores the following data to gain research insights:

1. **Outcomes:** Information about likes, reposts, replies, and quotes for each post. These types of outcomes have been widely used to study information contagion and cascade behavior in the research community [325, 326]

2. **Images:** Image data, along with the linguistic content of the posts can be used to study multi-modal communication aspects [327]

3. **Emojis:** As emoji use has increased on social media platforms, the need to analyze emoji use and how it correlates with language use has also grown [328].

4. **Conversation Threads:** This functionality allows analyses of conversations, similar to those made possible from Reddit discussions [329]

5. **Bridge Users:** Posts can become available across groups only when a bridge user who belongs to two groups interacts with them. This allows the controlled study of information contagion [330].

The conversation structure for the data collected is shown in Figure 3.2. Original posts form level 0 of the hierarchy. Any posts which are replies, reposts, or quotes on level 0 become level 1. Replies on level 1 become level 2, and so on - forming the tree of posts belonging to a conversation thread within the multiple co-occurring

conversations on the platform. In Community Connect, we utilize three main alphanumerically encoded identifiers - the "post_id", "parent_id" and "conversation_id". The "post_id" is a unique ID for each post. The "parent_id" is null for level 0, but stores the "post_id" of the parent of each post level 1 onward. The "conversation_id" is the same as the "post_id" for level 0, and each interaction on level 0 carries the same "conversation_id" as the parent. Thus, the "conversation_id" allows us to separate each conversation thread, and the "parent_id" allows us to reconstruct it by looking at the parent.

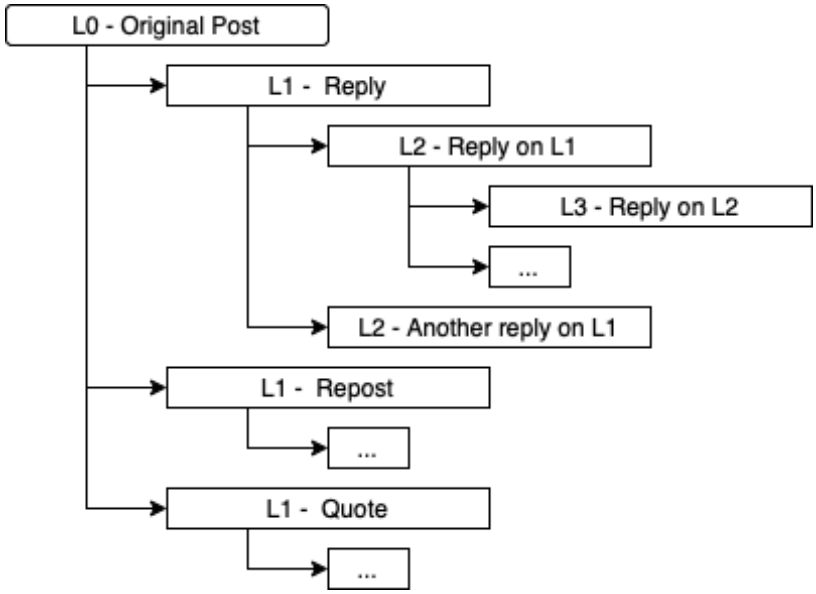


Figure 3.2: Thread structure of various kinds of posts via data captured in Community Connect. These could be potentially carry different topics within a substructure.

The ability to extract the conversational structure easily from all data captured on the platform allows for creating datasets that are rich not only in content, but also in the necessary metadata for solving tasks such as disentanglement and thread management. These properties are crucial for conversational AI systems to understand the ongoing conversation, especially since MPC often display multi-turn utterances and a model would need to understand the context of such a conversation via disentanglement to participate meaningfully [176]. Data from Community Connect also

contains metadata of the various threads that can arise within a conversation - not just the direct replies as seen in Figure 3.2 "L1 - Reply", but also spin-off conversations such as those in "L1 - Repost" and "L1 - Quote". Since Community Connect has similar affordances to Twitter, "reposts" can be used to express a direct and strong agreement, and "quotes" can be used to express an opinion on the original post.

Thus we have been able to utilize Community Connect in constructing feature-rich datasets for MPC modeling. The dataset creation process towards MPC modeling is described further in Section 3.5.

Table 3.2: Data collection statistics - Race is white (W) or minority (M); Gender is Female (F), Male (M), Other (O); Leaning is conservative (C), liberal (L), Independent (I). These categories have been crudely simplified for 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

3.5 Dataset creation for MPC modeling

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 to ensure diversity in the topics being discussed, following guidelines listed in [182] towards dataset creation, and make sure to remove personally identifiable information (PII) before utilizing the dataset.

3.5.1 User Experiment Setup

Much of our data collection efforts were underway during the COVID-19 pandemic, and our efforts to ensure equitable distributions in our participant pool were difficult (Table 3.2). Moreover, many conversations on the platform tended to focus around this topic. A larger team comprising of interdisciplinary researchers was involved to ensure good participation on the platform that reflected the behaviors observed in

```
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
vaccinations 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 3.3: An example of the zero-shot prompt for `flan-t5-xxl` to generate behavior annotations for each user.

emotional firestorms.

Utilizing Community Connect [153], we construct a structured social network, with roughly 15 sub-groups within the network. Each group is designed such that it is either heterogeneous or homogeneous - heterogeneous groups have a good mix of liberal and conservative leaning users (50-50), whereas homogeneous groups have an overall liberal or conservative leaning (80-20). The social network is then connected via bridge users, which connect groups in differing ways, such as connecting a heterogeneous group to a homogeneous liberal leaning group. Qualitative findings of this study are a separate effort and being conducted along with the interdisciplinary team.

To ensure the privacy of our participants, all participants are assigned fictitious names generated with a random name generator. We also assign fictitious email IDs to each participant, based on these generated names, via which they can log in to the platform. The password is selected by the participants themselves, which ensures that only that participant can login to their account. Once our dataset collection is complete, the personas are assigned to each participant, and any remaining identifiable

information is removed from the dataset.

Utilizing a mock social media platform allows us to gather the data with explicit user consent, and allows us to gather information about their race, gender, and leanings without PII, which enabled us to create a metadata-rich resource towards multi-party conversation modeling.

3.5.2 Social Media Behavior Categories and Annotation Methodology

One of the motivations for our study was to study how providing persona inputs can allow for response generation 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 [331].

Table 3.3: User behavior annotation statistics - first generated automatically using `flan-t5-xxl`, then manually checked for accuracy, and corrected.

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

Spectators are defined as participants who prefer to observe emotional conversations unfolding. They 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." Unlike the next two groups they are much less likely to seem to even consider that social media is 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. They prefer to discuss difficult topics but mainly share content they find positive, unifying, or productive. *Suppressors* suppress overly emotional content on social media during a firestorm. They view 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 a zero-shot prompting strategy to obtain automatic annotations for the typical behavior of each user based on the conversations they participate in. The `flan-t5-xxl` model [332] is utilized for generating the annotations since it has been shown to be effective for instruction based classification tasks [333]. The zero shot prompt consists of an instruction, an example of which is provided in Figure 3.3. We experiment with variations of the prompt, most notably trying to generate annotations for all users in the conversation at once, but find that the model performs more deterministically when the users are explicitly mentioned in the prompt, making it easier to collect all annotations.

Once the annotations are generated for each user in the dataset, they are manually checked for accuracy by 2 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 3.3. It is worth noting that Expressors form the 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.

3.5.3 Dataset format for MPC modeling

The fields in Community Connect allow us to compute and format our dataset to resemble the format of the Ubuntu Chat corpus [79] as described by Hu et al [174].

Table 3.4: Input fields required for modeling, and how they were derived from the dataset collected via Community Connect. We introduce the persona fields using survey outcomes and manually corrected annotations generated using `flan-t5-xxl`.

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 and manually corrected annotation generated using `flan-t5-xxl`
ctx_adr_persona	List of personas for each addressee in context	Compiled from user survey and manually corrected annotation generated using `flan-t5-xxl`
ans_spk_persona	Speaker persona of utterance to be generated	Compiled from user survey and manually corrected annotation generated using `flan-t5-xxl`
ans_adr_persona	Addressee persona of utterance to be generated	Compiled from user survey and manually corrected annotation generated using `flan-t5-xxl`

Table 3.5.3 lists each of the fields utilized for conversation modeling [174, 12, 16], and describes how each of these fields was computed from the data storage tables utilized in Community Connect. The persona related fields are introduced by us, and these fields store the persona properties for each utterance's speaker and addressee in a similar format to how the speakers and addressees themselves are stored.

3.6 Discussion

The collection, creation and annotation of this dataset has been a labor-intensive project, and has yielded a metadata-rich resource. The statistics related to the content of the dataset are presented in Table 3.5. 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. This dataset is available upon request, with access contingent upon approval.

Table 3.5: Dataset statistics for data collected via Community Connect. Final formatted dataset contains aggregated data from all 3 experiments towards persona-aware response generation.

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

We also observe that persona-based response generation models towards MPC modeling utilize two-party conversations in research owing to the lack of datasets that provide persona features for conversations [334]. However, with the data collected with Community Connect, each participant goes through an entry and exit survey which

not only provides Informed Consent but also records their political leanings. Knowing these properties, we are able to construct different naturally occurring property-based groups (ex, homogeneous liberal groups with mostly left-leaning members, and heterogeneous groups with a balance of left- and right-leaning members). This allows us to have significant metadata toward how different people can behave in different environments online, and can help us understand how people *audience design* when responding in groups with known properties, which is a part of future work currently being conducted along with an interdisciplinary team.

These experiments and the resulting dataset further showcase the capabilities of Community Connect, and how it can contribute meaningfully towards data collection for multi-party conversation modeling. We make Community Connect available online, and also provide the scripts to process the datasets collected via Community Connect to help with the creation of future feature-rich datasets.

CHAPTER 4: Persona aware Response Generation

4.1 Introduction

Conversational AI aimed toward generating utterances in a conversation has been a focal point of academic as well as industrial research for a long time. Both retrieval and generative strategies have been introduced in making this happen [335], with the conversational setting being a bot and a user, with the bot giving a response to the user's prompt. However, this prior work is limited to two-party conversations. Recently, there has been research focusing on MPC for response selection and generation, detailed in Section 2.5.2. Considering the added challenges towards MPC modeling (Section 2.2), there is a further need to explore the contributions that newer architectures such as Transformers and LLMs can make toward improving the response selection and generation capabilities of MPC systems.

Additionally, engagement with conversational partners is one of the most important traits for conversational agents [336]. Apart from being engaging, other traits that are considered desirable for conversational agents include being consistent [31, 337] and generating content that is affectively appropriate for the given situational context [338]. Consistency in conversational agents has been addressed by incorporating personality [31, 337]. We therefore propose utilizing persona attributes displayed by the participants in a conversation towards response generation in MPC.

To this end, we study how user attributes such as race, gender, and behavior type on social media (Expressors, Suppressors, Avoiders, and Spectators - refer to Section 3.5.2) might contribute towards generating responses that are more relevant, helping the conversation along by participating in keeping with the persona attributes of the responding user. We utilize HeterMPC [16] towards this task (Section 4.3), and

investigate the performance with and without persona attributes. We evaluate the model utilizing automatic and human evaluation strategies. We follow automatic evaluation strategies similar to HeterMPC [16], adding human evaluations not just for checking 1) relevance, 2) fluency and 3) informativeness of the response according 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. Response generation models which take persona attributes into account for modeling in two different ways - as an encoding concatenated with the utterance, and modeled as nodes connected to the utterance in the graph

2. Human evaluation scores towards a better understanding of the model's performance for the multi-party conversation response generation task

4.2 Related Work

Investigations of how persona attributes might affect generation are very limited for MPC modeling. Thus, we first begin with related work towards response generation in MPC modeling, then focus on persona-level datasets and existing work in persona MPC modeling, and lastly we discuss the factors which motivated our model evaluation strategy. We limit discussion to research focused solely on MPC modeling - while the substantial work on persona related dialogue modeling is a motivation towards our goal, we focus on the MPC modeling aspect since it is more central to our goal.

Response Generation. Zhang et al [339] propose a tree-based model frame for structure-aware group conversations, organizing the group conversation as a tree with different branches involving multiple conversation threads. They utilize hierarchical

encodings with the Seq2Seq encoder-decoder model [34] implemented with GRUs [340]. They outperform approaches evaluated on two-party dialogue modeling with the Ubuntu Corpus [79]. Liu et al [181] 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 [31], and VHRED [341] on their dataset. Hu et al [174] 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 [342] (of which VHRED [341] is the successor), both trained towards two-party dialogue modeling. Recently, [16] present HeterMPC, a heterogeneous graph-based neural network 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. The model architecture uses Transformers [343], with evaluation over the Ubuntu Corpus which outperform Seq2Seq [34], Transformers, and GSN by a statistically significant margin. We base our model architecture on HeterMPC owing to its performance as compared to previously proposed approaches to model persona-level attributes (details in Section 4.3).

Modeling persona attributes. Persona related research in MPC modeling is limited, with PersonaTKG [344] being the only model proposed towards this end 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 utilize utterance and persona nodes, 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++ [345], a dataset created by the authors, and compared with Seq2Seq, DialogueGCN [346], SIRNN [9], and PostKS [347], 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. Their modeling task includes persona tags which are both detailed and laconic, with input consisting of persona descriptions for the speakers. The model focuses on predicting the addressee and generating a contextual response, whereas our aim is to provide speaker and addressee information as input towards response generation. Additionally, PersonaTKG utilizes 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 [16]. A closer look at the dataset they utilize (HLA-Chat++) also reveals that extracting relevant fields towards modeling is not straightforward (refer to Table 3.4 for required fields), and scripts for calculating 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.

Our search for MPC with persona-level attributes thus continues with a recent survey on this topic [182], which lists two corpora which have the data we require for modeling. One of these is the FriendsPersona corpus [124] and the other is the TEAMS entrainment corpus [55]. FriendsPersona does not provide user level personas, and TEAMS does not provide the explicit speakers and addressees of each utterance in the conversation. Another corpus which could be useful is the PersonaChat corpus

[337], however PersonaChat also does not have conversation level data, with defined speakers and addressees, and personas for each utterances. Our modeling task requires mentioning speakers and addressees, and each of their personas which are a constant property of the user. This property is not available for these datasets - another reason we chose to collect and create our own dataset. Further details on dataset collection and persona attributes are discussed in Section 3.5.

4.3 Response Generation Model

Owing to the capabilities of HeterMPC [16] in regards to 1) response generation from scratch, 2) modeling the MPC as a conversation graph and supporting the generation of an utterance anywhere in the graph [174], and 3) support for utilizing the attention mechanism and Transformer architecture for modeling [348], we utilize HeterMPC as the base model. We add parameters which allow us to model personas, and discuss the implications (Section 4.3.2). The format of the data includes providing the speakers, addressees, and utterances as input to the model. The output of the model is a response which is generated from scratch (a generation task, not a selection task). Refer to Chapter 3 Table 3.4 for details of how we utilize the fields collected via Community Connect and format them for modeling.

4.3.1 Background

Our main task is response generation for MPC modeling for informal conversations which take place on social media platforms. We utilize HeterMPC [16] for this task, owing to its suitability towards modeling aspects of multi-party conversations for our task and availability on Github [349]. We also opt to utilize HeterMPC given its performance towards modeling the Ubuntu Chat corpus [79] - a dataset with properties similar to our persona dataset in terms of informal, asynchronous conversations, where the speakers and addressee information is known. Formally, given the conversation history and interlocutor information, we aim to generate $\bar{r}$:

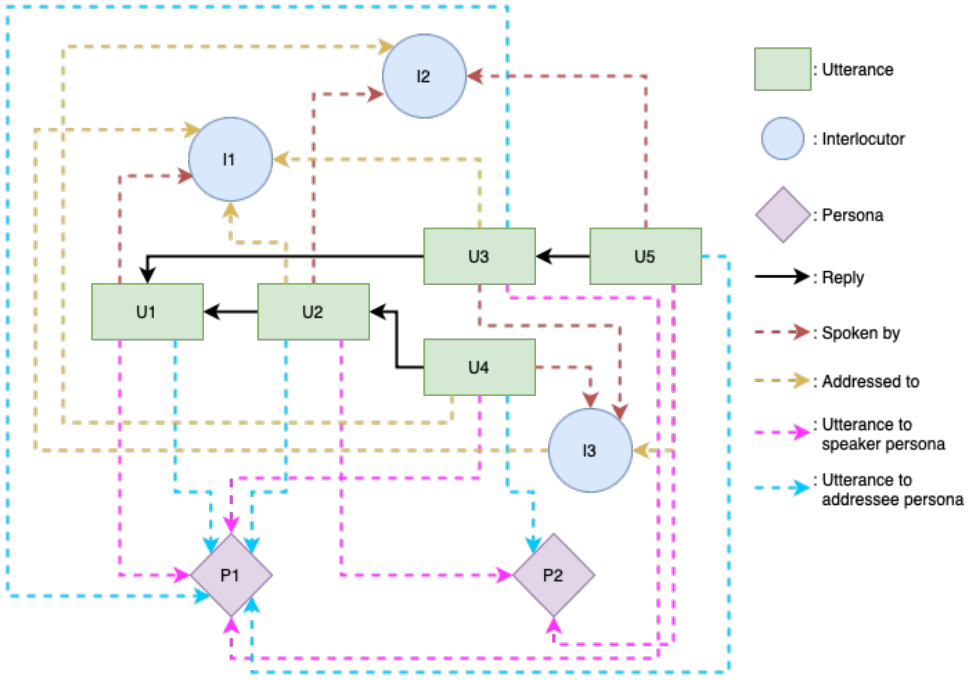


Figure 4.1: An example of the structure of the conversation graph, as modeled in HeterMPC [16] - modified to showcase the entire conversation structure and persona attributes.

$$\bar{r} = \underset{r}{\text{argmax}}\, logP(r|\mathcal{G})$$

$$= \underset{r}{\text{argmax}} \sum_{k=1}^{|r|} logP(r_k|\mathcal{G}r_{<k}) \tag{4.1}$$

HeterMPC utilizes a graph-like structure to model conversation flow using the DGL library [350]. An example of this is shown in Figure 4.1. The network utilizes a heterogeneous graph with 2 types of nodes, one representing utterances and the other representing interlocutors. The relationships between these are represented by 6 types of edges (only 3 are shown in the figure for brevity) - *reply*, *spoken-by*, *addressed-to*, and reverse direction edges for each (*replied-by*, *spoken-to*, *addressed-by*). The calculations discussed below are showcased in Figures 4.3 and 4.4 which have been modified to include persona information - details of which are in Section 4.3.2.

4.3.1.1 Node Initialization

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 [351, 352]. If an utterance represented by node n replies another utterance represented by node m, 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.

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 [250]. The Transformer architecture is utilized to encode and learn contextual representations [343]. The calculation for an utterance at the l-th Transformer layer is denoted as:

$$\mathbf{H}_m^{l+1} = TransformerEncoder(\mathbf{H}_m^l), \tag{4.2}$$

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 [178]. Since the order of each interlocutor is determined relative to their utterance in a given conversation, there is no need to learn representations separately for each user in the conversation, and the

order-based learning can be used across train, validation, and test sets.

4.3.1.2 Node Updating

Node representations are updated by feeding them into the built graph for absorbing context information [353, 354, 355]. Heterogeneous attention weights between connected nodes are calculated and messages are passed over the graph in a node-edge-type-dependent manner, inspired by introducing parameters to maximize feature distribution differences for modeling heterogeneity [356, 357, 358, 348]. Node-type-dependent feed-forward networks (FFNs) followed by a residual connection [359] are employed to aggregate information. A Transformer layer is placed specifically for the utterance nodes to update utterance encodings using graph information. L_2 denotes the number of iterations for updating both utterance and interlocutor nodes.

Heterogeneous Attention. Owing to the heterogeneous nature of the graph, node-edge-type-dependent linear transformations are applied to node representations before attention calculation so that the two types of nodes share similar feature distributions [357, 348]. 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:

$$\mathbf{k}^l(s) = \mathbf{h}_s^l \mathbf{W}^K_{\tau(s)} + \mathbf{b}^K_{\tau(s)}, \tag{4.3}$$

$$\mathbf{q}^l(s) = \mathbf{h}_s^l \mathbf{W}^Q_{\tau(t)} + \mathbf{b}^Q_{\tau(t)}, \tag{4.4}$$

$$w^l(s, e, t) = \mathbf{k}(s) \mathbf{W}^{ATT}_{e_{s,t}} \mathbf{q}(t) \frac{\mu_{e_{s,t}}}{\sqrt{d}}, \tag{4.5}$$

where $\tau(s), \tau(t) \in \{UTR, ITR\}$ distinguish utterance (UTR) and interlocutor (ITR) nodes. Eqs. 4.3 and 4.4 are node-type-dependent linear transformations. Eq. 4.5 contains an edge-type-dependent linear projection $\mathbf{W}^{ATT}_{e_{s,t}}$ 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:

$$\mathbf{v}^l(s) = \mathbf{h}_s^l \mathbf{W}_{\tau(s)}^V + \mathbf{b}_{\tau(s)}^V, \tag{4.6}$$

$$\bar{\mathbf{v}}^l(s) = \mathbf{v}^l(s) \mathbf{W}_{e_{s,t}}^{MSG}, \tag{4.7}$$

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:

$$\bar{\mathbf{h}}_t^l = \sum_{s \in S(t)} softmax(w^l(s, e, t)) \bar{\mathbf{v}}^l(s), \tag{4.8}$$

where $S(t)$ denotes the set of source nodes. The summarized message $\bar{\mathbf{h}}_t^l$ is aggregated with the original node representation $\mathbf{h}_t^l$ [359] as:

$$\mathbf{h}_t^{l+1} = FFN_{\tau(t)}(\bar{\mathbf{h}}_t^l) + \mathbf{h}_t^l \tag{4.9}$$

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:

$$\hat{\mathbf{h}}_t^{l+1} = [\mathbf{h}_t^l; \mathbf{h}_t^{l+1}] \mathbf{W}_{com} + \mathbf{b}_{com}, \tag{4.10}$$

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

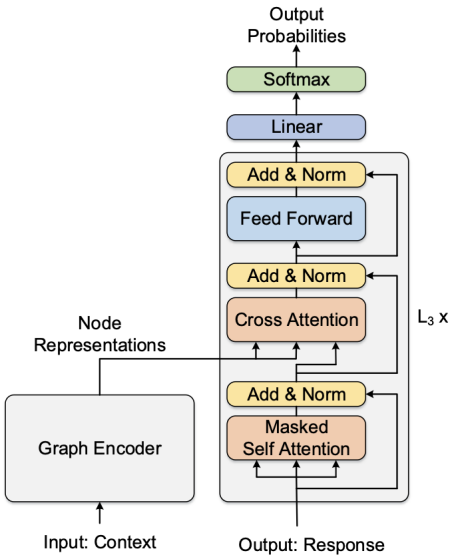


Figure 4.2: The decoder architecture of HeterMPC [16].

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 4.2). 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.

4.3.2 PersonaHeterMPC Model Architecture

We experiment with two different ways of modeling persona information towards response generation. First, we study the effect of concatenating the speaker and addressee persona attributes to each utterance. Next, we study how including the speaker and addressee personas as graph nodes, and introducing edges to the utterance nodes, affects response generation.

4.3.2.1 Together with utterance encodings

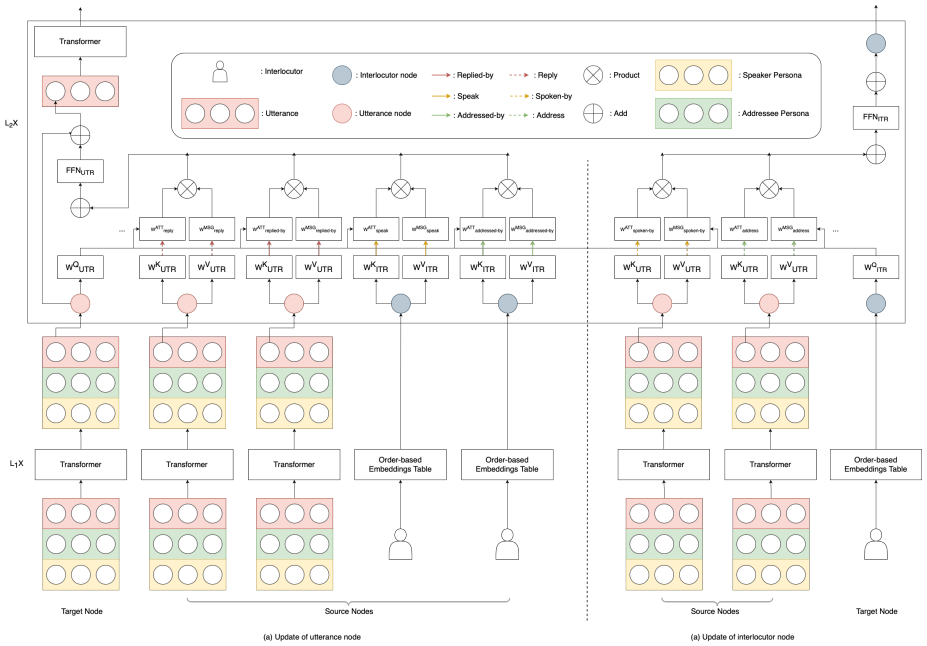


Figure 4.3: PersonaHeterMPC$_{\text{concat}}$, derived from the HeterMPC model [16]. We include persona attributes concatenated to utterance encodings explained in Section 4.3.2.1. The colors for the graph relations are coded similar to the relations showcased in Figure 4.1.

We utilize HeterMPC to enable support for persona modeling towards response generation in keeping with the speaker and addressee personas given as input for the utterance to be generated. The updates made to the architecture are presented graphically in Figure 4.3. The most notable changes are detailed below. Formally, the response generation task in HeterMPC includes modeling the context utterances $U = \{U_1, U_2, ..., U_m\}$, where m is the number of utterances, speakers and addressees $I = \{I_1, I_2, ..., I_n\}$ where n is the number of interlocutor, and modeling the relationships between each utterance U_k and interlocutors $I_k spk$ and $I_k adr$ as a graph. Our task is to also model personas $P = \{P_1, P_2, ..., P_n\}$ corresponding to each interlocutor in the conversation. The goal is to generate response $Y = y_1 y_2 ... y_s$ based on context and persona set, where y_i denotes the word generated in each step.

Persona encodings. Both speaker and addressee personas are provided as inputs

along with the response generation task. The beginning and ending of the personas are marked with the [CLS] and [SEP] tokens, to indicate separate sequences. The encoding for the personas is then padded with the [PAD] token to fit the maximum persona length hyperparameter - similar to how the utterances are encoded towards node initialization.

Encoder & Decoder Inputs. The input encodings for the encoder contain the speaker persona encoding, addressee persona encoding, and the utterances 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 decoder attention mask is updated to reflect this change, so that the decoder will attend to the personas while generating responses. 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 loss for training the response generation model.

4.3.2.2 Modeled with new graph nodes and edges

We also experiment with creating persona graph nodes and edges that model the relationships of speaker and addressee personas to an utterance. In our case, given a set of M utterances, I interlocutors, and K overall sets of possible personas, a heterogeneous graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ is constructed. Specifically, $\mathcal{V}$ becomes a set of $M + I + K$ nodes. $\mathcal{E} = \{e_{p,q}\}_{p,q=1}^{M+I}$ is a set of directed edges. Each edge $e_{p,q}$ describes the connection from node p to node q. We introduce four new meta-relations for persona connections, 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

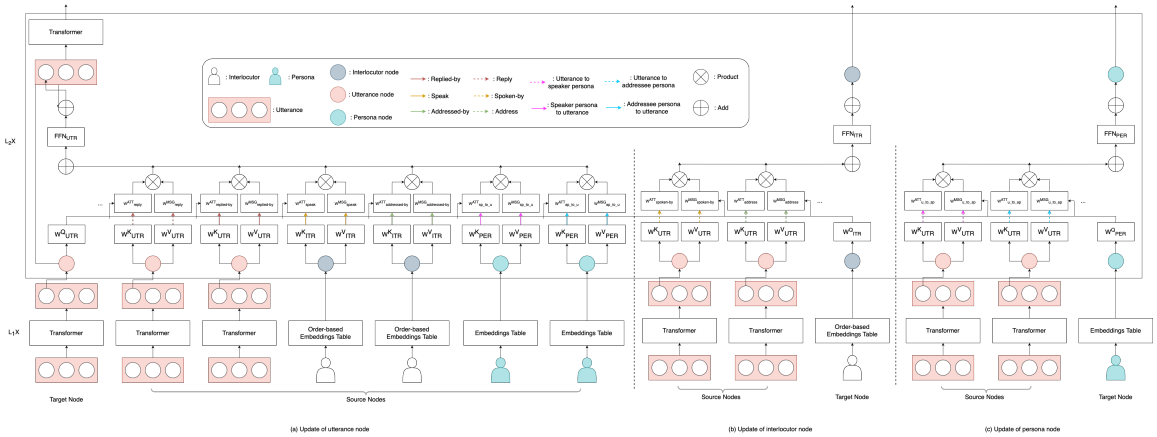


Figure 4.4: PersonaHeterMPC$_{\text{graph}}$, derived from the HeterMPC model [16]. We include persona attributes concatenated to utterance encodings explained in Section 4.3.2.2. The colors for the graph relations are coded similar to the relations showcased in Figure 4.1.

utterance represented by node n is spoken to an interlocutor whose persona is represented by node a, $e_{n,a} = \textit{utt-to-adr-persona}$ and $e_{a,n} = \textit{adr-persona-to-utt}$.

Node Initialization. The node initialization for utterance and interlocutor nodes remains the same as in HeterMPC. For persona nodes, node initialization differs than that of both utterances (modeled with Transformers encodings with `bert-base-uncased` [250]) and interlocutors (indexed according to their speaking order, the embedding vector is derived by looking up an order-based interlocutor embedding table). Since our aim is to study how different speaker and addressee persona properties affect response generation, the persona nodes are indexed globally over the entire dataset. They are then initialized with a global lookup table, and modeled with an embedding vector calculated on the basis of this value.

Node Updating. The overall model is shown graphically in Figure 4.4, nodes are updated by feeding them into the built graph for absorbing context information [353, 354, 355]. As with HeterMPC, heterogeneous attention weights between connected nodes are calculates and passed over the graph in a node-edge-type- dependent manner [356, 357, 358, 348]. To aggregate information after learning source to target node information, a node-type-dependent feed-forward network (FFN) [359] followed by a

residual connection is employed. A Transformer is placed for utterance node to let each token in an utterance have access to the information from other utterances. L_2 denotes the number of iterations for updating both utterance and interlocutor nodes.

We introduce a new node type, and thus follow the strategy of node-type-dependent linear transformations before attention calculation so that the nodes share similar feature distributions [357, 348]. The attention weights are then learnt similar to equations 4.3, 4.4 and 4.5; with message passing as defined in 4.6 and 4.7; and aggregation as defined in equations 4.8, 4.9 and 4.10; but for all 3 node types and 10 edge types.

4.4 Experiments

4.4.1 Response Generation Experiments

Much of the training hyperparameters were set similar to those of HeterMPC$_{\text{BERT}}$, utilizing `bert-base-uncased` pre-trained weights [360], optimization with AdamW [361], 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.

4.4.2 Other strategies studied for modeling personas

We experiment with a few other strategies while training together with utterance encodings - we experiment with (1) descriptive persona attributes, (2) using special tokens such as `[BOSP]`, `[EOSP]`, `[BOAP]` and `[EOAP]` to demarcate the persona sequences, and (3) randomly oversampling the dataset by a factor of $5, 10$ and 20.

We find that the best strategy overall for modeling together with utterance encodings is the descriptive personas. 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.*" The descriptions for the user behavior are similar to their definitions (refer to Chapter 3 Section 3.5.2). We report evaluation metrics with this strategy for modeling with utterance encodings in Section 4.5.

Using persona special tokens (`[BOSP], [EOSP]` for beginning and end of speaker persona, `[BOAP], [EOAP]` for beginning and end of addressee persona), we find that generations rank well on automatic evaluations. However, on closer look, we see that many generations begin either in the middle of a sentence or with a period, and are incoherent. We observe a similar issue with oversampling, and thus we do not pursue generations with these experiments (2 and 3) further.

4.5 Evaluation and Results

To support comparisons in future work, we follow the evaluation strategies detailed in HeterMPC [16]. Similar to previous work [174], we utilize the COCO evaluation package [362] 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

Table 4.1: Automatic evaluations for PersonaHeterMPC (PHMPC) compared to HeterMPC (HMPC) with different generation hyperparameters top_p and top_k - best values for each are in bold text.

Model	(top_p, top_k)	Metrics					
		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**
$\text{PHMPC}_\text{concat}$	(0.9, 5)	**13.118**	4.740	2.066	1.121	4.960	6.979
$\text{PHMPC}_\text{graph}$	(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**
$\text{PHMPC}_\text{concat}$	(0.5, 5)	11.305	4.358	2.068	1.285	4.594	6.574
$\text{PHMPC}_\text{graph}$	(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**
$\text{PHMPC}_\text{concat}$	(0.9, 10)	**12.085**	4.125	1.293	0.468	4.452	6.420
$\text{PHMPC}_\text{graph}$	(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**
$\text{PHMPC}_\text{concat}$	(0.5, 10)	10.533	3.809	1.616	0.961	4.509	6.678
$\text{PHMPC}_\text{graph}$	(0.5, 10)	**12.473**	**5.566**	**3.510**	**2.725**	**5.120**	8.733

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.

4.5.1 Automatic Evaluations

We present the results for three main response generation experiments in Table 4.1 - (1) the original HeterMPC model without persona information, (2) persona information modeled along with utterance encodings - $\text{PersonaHeterMPC}_\text{concat}$, and (3) persona information modeled as graph nodes with edges connected to utterance nodes - $\text{PersonaHeterMPC}_\text{graph}$.

We find that $\text{PersonaHeterMPC}_\text{graph}$ 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, upon a closer look, we find that many generations in these hyperparameters end up being NaNs (around 14%). In comparison, most generations for the hyperparameters we report in Table 4.1 produce NaNs in a much smaller range (about

5% to 7%). Thus, we include the generations obtained from these hyperparameters combinations. Additionally, we recognize that these generations might be affected by responses being images or gifs instead of text, and thus multimodal modeling for multi-party conversations is a direction for our future work.

4.5.2 Human Evaluations

We conduct human evaluations on the (1) base dataset, and generations by (2) HeterMPC, (3) PersonaHeterMPC$_{\text{concat}}$ and (4) PersonaHeterMPC$_{\text{graph}}$ towards 6 categories. In keeping with HeterMPC, we evaluate for relevance, fluency, and informativeness. We additionally conduct human evaluations towards understanding initiative-taking, thread relevance, and persona relevance. These metrics are defined as:

1. *Initiative-taking:* We first ask annotators to rate whether initiative was needed at the index where the response should go - for the reply to index of the response, would it help to to say something in the response which helps the conversation move forward? Then, we ask annotators to rate whether initiative was taken by the response, and whether it helped move the conversation forward.

2. *Thread relevance:* We ask annotators to rate whether the response makes sense as part of the conversation thread at the graph node where it should go.

3. *Persona relevance:* We ask for ratings for whether the response is suitable considering the speaker and addressee personas in regards to user behavior.

We report the average ratings given by two annotators in Table 4.2. 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.

Table 4.2: Human evaluation scores (averaged) for evaluating ground truth (Human), HeterMPC (HMPC) and PersonaHeterMPC (PHMPC) with utterance encodings and graph based modeling.

Models	Out of	Human	HMPC	$\text{PHMPC}_{\text{concat}}$	$\text{PHMPC}_{\text{graph}}$
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
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
Conversation-level avg	3	1.466	0.600	0.400	0.833

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.

4.5.3 Case Studies

We conduct two case studies, with one conversation less politically charged (Table 4.3) and one more politically charged (Table 4.4). In both conversations, we find that there are issues with fluency and grammar especially for $\text{PersonaHeterMPC}_{\text{concat}}$. Both HeterMPC and $\text{PersonaHeterMPC}_{\text{graph}}$ perform better in this regard.

We also change the speaker persona in regards to the social media behavior of the intended response to check whether the responses generated by $\text{PersonaHeterMPC}_{\text{graph}}$ perform well in keeping with it. We see that the generations vary with the speaker persona, and reflect the speaker persona well. Especially in Table 4.4, we can see that this property of $\text{PersonaHeterMPC}_{\text{graph}}$ is reflected in keeping with the political and thus emotional charge of the conversation as well as the speaker persona.

Thus we find that our work provides a baseline for modeling persona-aware response generation towards open domain MPC modeling. There are limitations and future

Table 4.3: Case study 1 for comparing ground truth, and generated responses by HeterMPC (HMPC) & PersonaHeterMPC (PHMPC).

Speaker		Addressee		Utterance		
ID	Persona	ID	Persona	ID	Parent ID	Text
1	white female conservative avoider	-1	-	0	-	just finished midnight mass on netflix! amazing show!
2	white male conservative expressor	1	white female conservative avoider	1	0	haven't seen it. what is it about?
3	white female liberal expressor	1	white female conservative avoider	2	0	right? if you haven't watched mike flanagan's other netflix shows (haunting of hill house, haunting of bly manor) i recommend checking them out. he's so good at twisting human drama/tragedy with supernatural elements, and his shows are beautifully shot.
4	white female conservative expressor	1	white female conservative avoider	3	0	pray for our country while you sleep soundly in your fantasy land
5	white female conservative avoider	2	white male conservative expressor	4	1	it is about a small island and a church there and has some sci fi elements. i don't want to give away the plot twists.:) (**Human**)
5	white female conservative avoider	2	white male conservative expressor	4	1	i love this! (**HMPC**)
5	white female conservative avoider	2	white male conservative expressor	4	1	.'ll be a good. (**PHMPC$_{\textbf{concat}}$**)
5	white female conservative avoider	2	white male conservative expressor	4	1	it's really nice. (**PHMPC$_{\textbf{graph}}$**)
5	white female conservative expressor	2	white male conservative expressor	4	1	. (**PHMPC$_{\textbf{concat}}$**)
5	white female conservative expressor	2	white male conservative expressor	4	1	it's pretty scary. (**PHMPC$_{\textbf{graph}}$**)
5	white female conservative spectator	2	white male conservative expressor	4	1	is the time of the most of the time. (**PHMPC$_{\textbf{concat}}$**)
5	white female conservative spectator	2	white male conservative expressor	4	1	i agree (**PHMPC$_{\textbf{graph}}$**)
5	white female conservative suppressor	2	white male conservative expressor	4	1	.'t have to me. (**PHMPC$_{\textbf{concat}}$**)
5	white female conservative suppressor	2	white male conservative expressor	4	1	i don't have the right. (**PHMPC$_{\textbf{graph}}$**)

Table 4.4: Case study 2 for comparing ground truth, and generated responses by HeterMPC (HMPC) & PersonaHeterMPC (PHMPC).

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$_{\textbf{concat}}$**)
5	white male liberal spectator	1	white male independent expressor	4	1	it's not a good one! (**PHMPC$_{\textbf{graph}}$**)
5	white male liberal expressor	1	white male independent expressor	4	1	that's right.. (**PHMPC$_{\textbf{concat}}$**)
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$_{\textbf{graph}}$**)
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$_{\textbf{concat}}$**)
5	white male liberal suppressor	1	white male independent expressor	4	1	it's just a lot of the real. (**PHMPC$_{\textbf{graph}}$**)
5	white male liberal avoider	1	white male independent expressor	4	1	is the time to do. (**PHMPC$_{\textbf{concat}}$**)
5	white male liberal avoider	1	white male independent expressor	4	1	they are right! (**PHMPC$_{\textbf{graph}}$**)

work needed towards realizing this goal more meaningfully, which we discuss in Section 4.6.

4.6 Discussion

We introduce a dataset with persona-level attributes towards multi-party conversation response generation. It includes attributes towards race, gender, political leaning and social media behavior. We experiment with various strategies towards persona-aware response generation for MPC, and find that modeling personas into the conversation graph works best towards our task. Lastly, we conduct human evaluation in keeping with known shortcomings in MPC evaluation (Chapter 5), and introduce new albeit subjective methods towards gauging the effectiveness of persona-aware MPC response generation.

Through our work, we find that persona-aware generation could indeed help towards better MPC modeling. With multiple participants in the conversation, having more information about their behavior could support an MPC model in generating responses which are more relevant as it takes part in a conversation. Our work falls into type 2 (active-simulative) systems (Chapter 2 Section 2.2). One future direction in our work includes conducting research towards type 1 (active-additive) systems, and our case studies (Section 4.5.3) prove that this work could be extended towards this end by introducing new speakers that could participate in the same conversation.

There are many other future directions which could help further our aim towards studying persona-aware MPC response generation. One of these includes experiments with HeterMPC$_{\text{BART}}$. Another involves experimenting with the number of Transformer layers for initialization (L_1), the number of iterations for updating both utterance and interlocutor nodes (L_2), and the number of decoder layers (L_3). These experiments are computationally intensive, and thus require time and resources to complete, thus counting towards future work. They could potentially help mitigate the effect of the difference in dataset size (the Ubuntu Chat dataset is significantly

larger than our dataset - it is at least 50 times larger). Additionally, some data points in our dataset contain images and gifs instead of text (around 5%). There is a need to experiment with multimodal models towards modeling these conversations better.

These future directions also point to the limitations of our work. While we experiment with oversampling towards accounting for the difference in dataset sizes, there are other strategies which have not been investigated, such as utilizing cross validation when finding the best model towards generation. Another major limitation is that our experiments are conducted purely in English, and it is worth studying the effect of persona-aware modeling for multilingual data (along with multimodal modeling). We aim to conduct future studies towards overcoming these limitations.

CHAPTER 5: Evaluation for Multi Party Conversation Modeling

5.1 Introduction

There has been much discussion lately in the field of Natural Language Generation (NLG) focusing on the need for evaluation benchmarks and standards, as evidenced by the prolific literature focusing on the issues surrounding human evaluation [363, 364, 365, 366, 367], as well as recently proposed benchmarks [368, 369, 370, 371]. These are important and necessary debates - however, work has focused mainly on two-party dialogue systems. Multi-party dialogue (MPC) systems, which aim to model conversations between groups (>2 participants) have received less attention, especially in the area of evaluation. Additionally, while there is existing work towards modeling MPC, evaluation strategies are not consistent across existing literature, making it harder to place the progress of the field. In the context of multi-party dialogue (MPC), we discuss both automatic and human evaluation metrics used for evaluating the three main sub-tasks described in detail in Section 5.2.

Thus, we foreground the challenges faced by the presence of multiple participants in a conversation, and how this property affects the evaluation of systems that aim to model group conversations. We present an expansion to the integrated taxonomy (Table 5.1) [2]. We use [2] as a baseline owing to their extensive study of data-driven and theory-driven error analysis, and the empirical validation of the proposed integrated taxonomy drawn from both these error analysis paradigms [372, 373]. However, we find that the integrated taxonomy does not account specifically for the challenges faced by MPC modeling systems, and thus we propose an expansion specifically keeping these challenges in mind.

We then draw attention to specific shortcomings of evaluation metrics utilized in

existing work, such as the lack of consistent reporting within similar evaluation metrics (such as $Recall_n@k$), and the lack of public availability of the proposed methodologies, making it harder to place the progress of the field even if an evaluation benchmark is proposed. Thus, there is a severe gap toward a consistent evaluation framework in Multi-Party Dialogue (MPC) which needs to be addressed. Our main contributions include:

1. We propose an expanded taxonomy focusing on the specific challenges introduced by multi-party dialogue, or group conversations (such as the need to maintain speaker-specific context and recognize the proper addressees), and provide examples for each newly introduced category.

2. We synthesize evaluation measures currently used in MPC research and relate them to the expanded taxonomy introduced.

To study evaluation metrics in existing work, we surveyed over 338 research papers in the field of MPC [42]. We obtained the initial pool based on a keyword search for variations of "multi-party dialogue", with 258 papers focused on work in English, and most of them published at *CL, LREC, and related conferences. The papers included in this article include only those that (a) focus on the English language, (b) include *multiple speakers in the majority of conversations*, and (c) which focus on text-based approaches (thus excluding research that uses multi-modal cues towards the aforementioned sub-tasks). This paper does NOT focus on multilingual corpora, or approaches that solely focus on concepts such as speech recognition or synthesis. We also limit discussion to research published within the past decade for a more relevant understanding of the current progress in MPC modeling, and aim to build upon limited prior work in MPC evaluation, which we discuss further in Section 5.3.3. With this filtering, we find a total of 15 papers whose aim is one or more of the sub-tasks of Speaker Identification, Addressee Recognition, and Response Selection/Generation.

We first present an expanded taxonomy with error reporting drawn from the challenges presented in MPC [7] and [48], adding categories specifically relevant and important towards MPC evaluation to the taxonomy [2]. Next, we observe the evaluation metrics utilized in existing work in Section 5.4, whose error reporting strategies we relate to the proposed expanded taxonomy (Table 5.1) and note the lack of evaluation for important categories.

5.2 Overview of Challenges in MPC Evaluation

Evaluation for MPC has often focused on specific sub-tasks that are integral to the working of any conversational system participating in a group conversation. A lot of existing research focuses either on one or more of the sub-tasks: 1) *Speaker Identification* which concerns how an MPC chatbot is able to track the speakers for each utterance as well as predict who the next speaker could be, 2) *Response Selection* which concerns with the selecting the correct next utterance from a set of choices or *Response Generation* which concerns with generating the next utterance from scratch given the context of the conversation, and 3) *Addressee Recognition* which concerns with being able to find the addressee(s) for the next utterance. All Speaker Identification, Response Selection, and Addressee Recognition can be framed as classification tasks (evaluation would need to check whether the correct participant(s) were chosen from the group), whereas Response Generation requires evaluation metrics similar to response generation for two-party dialogue. Recently, systems trained towards jointly modeling one or more of the above tasks have been proposed, however as mentioned before the evaluation strategies lack consistency, and require further thought. While evaluating the classification could provide important indicators of the performance of the dialogue model itself, robust evaluation is needed to understand how well the system would perform in a real life setting. Some leading questions which venture into this challenge faced by MPC systems include:

1. Is the system able to maintain long-term context from all participants in the

group? Is the selected/generated response relevant to the prompt and the context of the MPC participants while being grounded in the ongoing conversation? (Pointing to the need for managing speaker information)

2. Is the system able to respond to every participant's prompt, whether implicitly or explicitly mentioned? Conversely, is it able to learn to not respond (yet remember for context) to the relevant utterances? (Pointing to the need for managing addressee information)

3. Does the system contribute towards making the conversation successful? This success could be attributed to either making the conversation easier for the group by providing information when needed, measuring the interactivity introduced by the presence of MPC dialogue systems, and helping the group achieve the objective which led to the conversation. (Pointing to the need for evaluating appropriate timing and thread management abilities)

Keeping these challenges in mind, we present an expansion of error reporting categories which would be the first step toward accounting for the performance of a system that operated in the multi-party conversation. We briefly summarize the error reporting taxonomy for dialogue agents [2], and then discuss how the expansion accounts for errors specific to multi-party dialogue in Section 5.3.

5.3 Expanded Taxonomy of Errors for Multi Party Conversation

Recently, an integrated taxonomy of errors in chat-oriented dialogue systems (Table 5.1) was introduced [2]. Their work focuses on responses given by a chatbot (conversing with one user) which could cause a breakdown in the conversation [372]. They empirically validate the resulting integrated taxonomy by asking the same annotators who annotated breakdowns to rate the breakdown for each error category [373]. While the resulting taxonomy is quite exhaustive, we find that it does not account for challenges specific to MPC, such as the need to know whether the user is able to

Table 5.1: Integrated taxonomy for errors in chat-oriented dialogue systems [2]. We expand the taxonomy to include errors specific to MPC - extensions are italicized. The numbering is assigned serially and used in text to refer to discussions surrounding the specific error.

	Violation of Form	**Violation of Content**
Utterance	(I1) Uninterpretable (I2) Grammatical error	(I3) Semantic error (I4) Wrong information
Response	(I5) Ignore question (I6) Ignore request (I7) Ignore proposal (I8) Ignore greeting	(I9) Ignore expectation *(I18) Forgot speaker* *(I19) Forgot addressee(s)*
Context	(I10) Unclear intention (I11) Topic transition error (I12) Lack of information	(I13) Self-contradiction (I14) Contradiction (I15) Repetition
Society	(I16) Lack of sociality	(I17) Lack of common sense
Participant	*(I20) Wrong speaker* *(I21) Wrong addressee(s)*	*(I22) Wrong thread response* *(I23) Inappropriately timed initiative*

attribute utterances to each participant correctly. Thus, we expand the taxonomy [2], focusing specifically on how the presence of multiple participants affects the possible errors which occur in a group conversation.

We elaborate on each error from an MPC point of view, providing examples demonstrating the need for further research. We draw from perspectives relating to the challenges presented for realizing the differences between two-party and multi-party dialogue evaluation [7, 48]. Specifically, we expand on Response-level errors (I18 and I19) which are affected by the speaker and addressee(s), and add a new dimension with Participant-level errors (I20, I21, and I22), which showcase errors from a participant's point of view. We include all these, italicized and highlighted, in Table 5.1, and include details for each error with examples in this section.

5.3.1 Response level Errors

This subsection focuses on response level errors, which apply to the semantic meaning of the complex information contained in responses in MPC.

5.3.1.1 Violation of Content

We maintain the definition presented earlier [2], and thus violation of content errors indicate that even though the surface form of the utterance may be appropriate, it could lead to confusion during the conversation.

(I18) Forgot speaker: The utterances made by a specific user are often ignored. This relates specifically to the challenge of Speaker Identification [7], and is an extremely important property for maintaining context in MPC, since it could create confusion for the system downstream if the utterance is referred to again and the user feels ignored. In the example below, the System (S) forgets the utterance made by User 1 (U1) at the beginning of the conversation. Failure to remember the correct speaker for an utterance could lead to critical downstream errors.

(1) U1: We need to consider factors A and B for making a decision in case X.
U2: Factor C would also be interesting and important to consider along with A and B.
S: U2 mentions factor C will be important to take into consideration for case X.

(I19) Forgot addressee(s): The system forgets to mention the correct addressee(s), relating to the Addressee Recognition challenge [7], and specifically forgets one or more addressees it should have mentioned. If the system was prompted by multiple speakers on a similar topic, but the system responded only to some, this counts as an error since it could make forgotten participants feel alienated from the conversation. In the example below the System (S) forgets to address User 2 (U2), although it should have included both U1 and U2.

(2) U1: We need to consider factors A and B for making a decision in case X.

U2: Factor C would also be interesting and important to consider along with A and B.

S: Thanks for bringing factors A, B and C up for case X, U1.

5.3.2 Participant level Errors

We introduce a new broad category of errors for MPC evaluation called Participant-level errors. The categories of errors introduced in this section stem from the inherently entangled nature of responses in MPC - a response contains not only the content and context of the utterance but also information about who the speaker(s) and addressee(s) of the utterance are. We elaborate further in this subsection.

5.3.2.1 Violation of Form

The violation of form at the participant level indicates that the system involves the wrong participant(s) in the current state, leading to confusion.

(I20) Wrong speaker: The system credits the wrong speaker for an utterance, also relating to Speaker Identification [7], but specifically mistaking one user for another. This can lead to very important mistakes, since different participants could express differing views which could be contrasting in nature. In the example, the System (S) credits the utterance to U1 instead of U2, which diminishes U2's contributions.

(3) U1: We need to consider factors A and B for making a decision in case X.
U2: Factor C would also be interesting and important to consider along with A and B.
S: U1 mentions factors C will be taken into consideration for case X.

(I21) Wrong addressee(s): The system mentions the wrong addressee when responding to a prompt, leading to miscommunication. Relating to the challenge of Addressee Recognition [7] as well, mentioning the wrong addressee could lead to a breakdown since this means the system was unable to maintain contextual information that it should have in the conversation. This could lead to a breakdown especially if the addressee who is mentioned does not wish to be mentioned/take part in the current conversation. In the example, the System (S) mentions the wrong addressee U1 instead of U2.

(4) U1: We need to consider factors A and B for making a decision in case X.
U2: Factor C would also be interesting and important to consider along with A and B.
S: Interesting insight on factor C, U1.

5.3.2.2 Violation of Content

A violation of content means that the system makes an error that might seem appropriate in the conversation, but is incorrectly placed, therefore leading to confusion.

(I22) Wrong thread response: MPC can have communication ongoing in multiple threads within the same conversation (Thread/Conversation Management [7]).

If the system talks about the wrong topic when participating in a different thread, this could lead to confusion and interrupt the desired flow of conversation. In the example below there exist two threads of conversation: one whose topic is sports (U1, U2, U3) and the other whose topic is movies (U4, U5). There are sub-groups of users within the conversation who are participating in different threads, and the System (S) makes an error by mentioning a topic in the wrong thread and sub-group.

(5) U1: This football season has been going great!
U2: I agree, for most teams anyway. Which one is your favorite?
U3: I prefer soccer instead. Is anyone here a soccer fan?
U4: I don't pay much attention to sports. My main hobby is movies!
U5: Yeah, and Knives Out was a great one!
S: I agree U5! The Rams are doing so well this year!

(I23) Inappropriately timed initiative: MPC systems need to figure out when to take the floor in a conversation without causing an abrupt change in the conversation. Secondly, while they could be prompted to speak, it is also important to take the lead to get a conversation started since participants could be yielding the floor to other participants. This relates specifically to the challenge of Initiative Management [7], since the system needs to learn when to take initiative and introduce new topics without which the conversation might come to a halt. In the example, the conversation flow is smoothly going on for fiction (U1, U2, and U3), but the System (S) interrupts with a contrasting topic.

(6) U1: I love documentaries and it's been great seeing so many come out in recent years.

U2: They do seem informative. I'm particularly interested in performative documentaries, they seem more personal.

U3: I also enjoy performative documentaries, like Supersize Me. Have you watched it, U2?

S: Does anyone here like fiction?

5.3.3 Takeaways

In recent research, we observe the prevalence of the aforementioned errors within MPC research. We notice how the need to account for multiple participants affects the response selection/generation pipeline for systems modeling MPC, and thus discuss error reporting in existing research in the section to highlight our observations. Since there is limited existing research in the field of MPC response selection/generation, we reserve experimental validation of the expanded taxonomy for future work. However, the first research paper of particular interest to this discussion [14, 15] propose evaluations for interactions between virtual multi-party systems and users: 1) User Satisfaction via rated survey questions (accounting for Response-level errors I5-I9, I18, & I19, Society-level errors I16 & I17, and Participant-level errors I20-I23); 2) Intended Task Completion via predefined task success and inter-rater reliability (accounting for I4 and I12); 3) Recognition Rate via classification F-score (accounting for I19 and I21); and 4) Response Appropriateness via a custom-defined scale (accounting for Context-level errors I10-I15 and I22-I23). This paper presents a great first step in evaluations for MPC systems that interact in the real world, and we hope to draw from such studies for future work (Section 5.5).

5.4 Inconsistency of Evaluation Metrics in Existing Research

Papers focusing on specific tasks within MPC have been observed to employ mostly automatic evaluation measures, with very few including human evaluations. Repeated observations within mainly two-party NLG evaluation have shown that automatic and human evaluations do not correlate well [374, 375, 376, 377, 378], leading to arguments about automatic evaluations being unsuitable for assessing linguistic properties [379]. Owing to these, [367] survey the field and present arguments towards how the inclusion of human evaluations gives a more complete picture of the performance of systems whose main purpose is to participate in human conversations. With research in MPC severely lacking this reporting, it is difficult to place the success of systems that have been proposed to perform well in real-world scenarios. Moreover, owing to the complex nature of group conversations, this lack of reporting exacerbates the effect towards understanding the progress of MPC. Thus, this section illustrates research focusing on the core task of MPC modeling, drawing attention to the evaluation strategies followed by them. We provide a brief synthesis of currently formalized tasks and relate the errors from the expanded taxonomy (Table 5.1).

5.4.1 Evaluation Metrics in Sub tasks

We organize this section by including sub-task focused discussions to get a clearer idea of the evaluations reported for each sub-task, and how these relate to the expanded taxonomy of errors. We start with the joint formalized task introduced [8] - Addressee Recognition and Response Selection, Section 5.4.1.1 - which is the one of the most consistent research area with regards to error reporting. We then focus specifically on Response Selection in Section 5.4.1.2, then moving to Response Generation in Section 5.4.1.3, and lastly Speaker Identification in Section 5.4.1.4. Lastly, we wrap up by discussing the overall takeaways in Section 5.4.2.

5.4.1.1 Addressee Recognition and Response Selection

The first formalization of the Addressee and Response Selection (ARS) [8] joint task included an input consisting of the (responding agent, context, candidate responses) and output consisting of the (addressee, response). The formalization is accompanied by a few models, the best being the Dual Encoder based RNN model (called Dynamic RNN), for which they evaluate accuracy over addressee selection (ADR) limited to the addressee of the last utterance, and response selection (RES), as well as a mix of both with addressee-response pair selection (ADR-RES). SI-RNN [9] then utilize the same framework for their evaluation, improving their model by including speaker embeddings. Who2Whom [10] focuses on identifying addressees within the same task, but for all utterances, also reporting accuracy (with n-grams, n=5, 10, and 15) and $Precision$@1. They also involve limited human evaluations, comparing the consistency between human and model outputs, along with significance tests. MPC-BERT [12] introduces pre-trained models and fine-tuning for downstream tasks within MPC systems. They follow the same evaluation strategy established earlier [9].

Thus most papers in this line of research focus on measuring errors towards I18, I19, I20, and I21, with some including human evaluations for a subjective understanding of the success of their models.

5.4.1.2 Response Selection

Topic-BERT [11] and SA-BERT [178] focus on response selection as a classification task, with two very similar frameworks. The main difference between the approaches is that Topic-BERT builds topic-sentence pairs as input, while SA-BERT instead builds speaker embeddings as input - both utilize the basic embeddings for BERT pre-training (segment, position, and token embeddings). Both report recall as defined by the response selection task proposed in DSTC-8 [304] sub-tasks 1 and 2, using $Recall_n@k$ for reporting recall for matching n available candidates to k best-matched

responses (the official leaderboard utilizes MRR and $Recall@10$ with $n = 100$). However, their is still no overlap in the evaluation results for response selection on DSTC-8 reported by both papers, with one [11] reporting $Recall@1, Recall@5, Recall@10$ and MRR (assuming all these are reported for $n = 100$ - only mentioned in Section 4.1 of the paper) which details the pre-training for Topic-BERT; and the other [178] reporting only $Recall_2@1$, $Recall_{10}@1$, $Recall_{10}@2$, and $Recall_{10}@5$, although they do mention $Recall_{100}@1$ once in Section 1. Both papers do however mention $Recall_{10}@1$, $Recall_{10}@2$, and $Recall_{10}@5$ for the Ubuntu V1 corpus, which does allow a partial comparison for results. Additionally, Topic-BERT [11] also report BLEU [380] and $Precision@n$ (n=1, 2, 3, 4) scores for incorrectly selected responses, checking the relevance of the Topic-BERT retrieved results.

The Thread-encoder model [176] also tackle response selection, with more focus on dialogue dependency to organize the conversation into contextually aware threads (built with Transformer based BERT-base, same as [11, 178]. They utilize similar data (Ubuntu V2 and DSTC-8), and report evaluations for response selection, reporting $hits@k$ (similar to $Recall@k$ as per the paper and ParlAI [381] metrics, $k = 1, 2, 5$), and MRR for Ubuntu V2 and $hits@k$ (similar to $Recall@k$, $k = 1, 5, 10, 50$) and MRR for DSTC-8 (with n=100).

Since most papers working on response selection essentially work on a classification task, naturally the reporting is limited to classification metrics. However, even research conducted around the same time, over the same task, reports different metrics with only partial overlaps which could be used to partially compare performance. However, we do not consider this evaluation to count towards any of the expanded taxonomy since none of the classification metrics specifically look for performance consciously in any of the dimensions included in the taxonomy - they just measure whether the system was able to choose the next utterance given the previous utterances and a possible list of the right next utterance. Breaking down the evaluation

into components presented in the taxonomy, i.e. measuring success keeping in mind the speaker, addressee, and content & context of the selected utterance would help understand the performance more robustly - like [11] report BLEU for the incorrect responses.

5.4.1.3 Response Generation

Response generation is formalized as an input consisting of previous utterances and an output consisting of the next utterance, and has been tackled with various methods [179, 181, 174] ([181] also specifically include the responding speaker and target addressee in the inputs and outputs). One [179] reports the BLEU-n (n based on n-grams, n = 1, 2, 3, 4) and METEOR [382] scores (mentioning that the evaluation could be supplemented); another [181] reports BLEU, ROUGE [383], noun mentions, and length of generated response, along with limited human evaluations for fluency, consistency, and informativeness; and the last [174] reports BLEU-n (n = 1, 2, 3, 4), METEOR, ROUGE-L (L for longest common subsequence), along with human evaluations for fluency, grammaticality, and rationality. Models utilizing the conversational structure [172, 174] towards response generation, with structured attention and Variational RNN, report the same automatic metrics BLEU-n (n = 1, 2, 3, 4), METEOR, ROUGE-L (L for longest common subsequence). They also find that they can perform speaker identification and addressee recognition without specifically training for these tasks.

Multi-role Interposition Dialogue System (MIDS) [245] tackle response generation along with speaker identification, proposing LSTMs to build an encoder, a contextual RNN, a speaker encoder, and a decoder. They report accuracy for speaker identification; and perplexity and loss for response generation.

Even with the majority of papers reporting the basic automated evaluation metrics most common for generation (BLEU, METEOR, ROUGE [367]), these are not always reported. Moreover, the aforementioned metrics show either weak or no correlation

with human judgments [384]. Human evaluations are also limited, although they do cover some of the most reported metrics (fluency, consistency, informativeness, grammaticality, rationality [367]). Most research thus covers major aspects of the expanded taxonomy, namely Utterance-level I1-I4, Context-level I10-I15, and Society-level I17. Some papers also report speaker identification and addressee recognition, accounting for I18, I19, and Participant-level I20-I23 with thread management.

5.4.1.4 Speaker Identification

Approaches using RNN and CNN to identify speakers with a sitcom dataset [169], report accuracy and F1 (+ F1 towards each participant and a confusion matrix to better analyze wrong predictions). Those using MLE, SVM, CNN, and LSTM architectures to model sitcom, finch and multibotwoz datasets [170] report accuracy. Both utilize a variety of features (such as surrounding utterance concatenation, agent and content information) with the models to improve predictions. They extend their work by integrating MLE, CNN, and FSA-based architectures, for multibotwoz, reporting accuracy [171].

Classification for speaker identification does help response selection and generation, counting towards errors I18 and I20 from the expanded taxonomy. However, it would be helpful to include more classification metrics (like the confusion matrix [169]) to allow for more robust evaluations.

5.4.2 Takeaways

It is imperative to observe the various kinds of evaluation metrics that have been used to evaluate different tasks within MPC. Most metrics reported are not consistent across the main task they focus on, sometimes even when they report performance on a shared task such as DSTC-8 [291]. It is important to note that these inconsistencies lead to confusion when it comes to looking for the current state-of-the-art systems, as well as for making important performance comparisons such as significance testing.

Additionally, we find that there is a 50-50% (8:7) division of the code in the papers being publicly available (if we include broken links, the unavailability goes up, but we count these as attempts to provide reproducible methods). This means that even with re-evaluation given a benchmark, there is a possibility that comparison across existing research will not be able to provide a full picture of the progress in each sub-task.

All these issues draw attention to the need for more shared tasks and robust benchmarks which report errors in a manner fitting the proposed taxonomy. We postulate that this would allow better comparisons across tasks, and overall performance towards building systems able to participate in MPC - although we reserve the evaluation of our proposed extensions to the taxonomy itself for future work. We aim to follow methods similar to the ones described by [373] to maintain the standards they set up for validation of error analysis.

5.5 Discussion

We have presented an expansion - which focuses specifically on errors important in multi-party dialogue - to the integrated taxonomy of errors [2]. We include examples for each newly introduced error in Section 5.3, and relate the errors to the MPC modeling challenges [7]. We then present inconsistencies in the evaluation strategies reported in existing research (Section 5.4), organized by the sub-tasks they focus on. We observe the difficulty in comparisons across the proposed methods owing to inconsistencies in error analysis. We also relate the reported errors to the expanded taxonomy, drawing parallels for overall comparison.

We observe how the challenges introduced by the presence of multiple participants affect the need for more robust evaluations (Section 5.3.3) which are capable of reporting how well the approach performs, and a preliminary discussion [14, 15] surrounding these errors, albeit more focused on interactions between virtual systems and users. We also find that even with defined tasks, inconsistencies could arise in reporting

errors (Section 5.4.2), leading to confusion when placing the progress of research in MPC.

We note that while our presented taxonomy is relevant to the errors reported in current literature, there is a need to evaluate their effectiveness empirically, which is the main limitation for this paper and proposed future work. Another big limitation of this work which is also a part of proposed future work is the formalization of the proposed expanded errors specific to MPC from this paper (Table 5.1), and the validation of the formalization toward a proposed benchmark. The first shared task DSTC-8 [291] focused on the response selection sub-task, however, there is a need for future shared tasks which account for all three sub-tasks (speaker identification, response selection/generation, and addressee recognition), and related sub-tasks (such as disentanglement, thread management, and coreference resolution).

We draw from our work in this area towards evaluating our persona-aware response generation model proposed in Chapter 4 (PersonaHeterMPC), specifically towards measuring initiative-taking (I23), and thread relevance (I22). We also ask for persona relevance scores from human annotators to gauge whether generated responses were in keeping with the expected behaviors of the speaker and addressee.

CHAPTER 6: Conclusions and Future Work

The field of multi-party conversational AI has yet to see significant strides as compared to two-party modeling. We have shown that many research gaps exist within this field, and aimed to fill some of these by (1) creating tools for corpora collection and utilizing these to create effective datasets towards persona-aware conversational modeling, (2) introducing a novel persona-aware multi-party response generation model, and (3) providing guidelines and introducing human evaluation metrics towards improving NLG evaluation for multi-party conversations.

We introduce the **Community Connect platform open-sourced on Github [153]** which acts as a mock social media network. It supports asynchronous communication among multiple participants who can log on from anywhere, and setting them into experimental groups which can be defined a the research team. It is built on the MEAN stack and can run on AWS, making it scalable in nature. **We utilize Community Connect in 3 distinct experiments (with around 150 participants in each) to collect a cumulative dataset of over 1800 conversations with more than 5 turns each.**

We observe that recent research focusing on utilizing Transformers towards response selection and generation tasks in MPC [178, 12, 16] utilize the Ubuntu Dialogue datasets, which were released in 2015 [79], since existing research utilizes the corpus as well [8, 9, 10, 174]. While the Ubuntu Dialogue corpus is important towards baseline benchmarking, it is also restrictive in the topics of conversations which are mainly related to the usage of the Ubuntu operating systems (OS) and questions about solving any errors that might arise while using the OS. Towards this end, we make the metadata-rich dataset we have been able to collect with Community Connect available

upon request, which are more reflective of everyday conversations. We provide the data in a format that can be **easily adopted toward MPC modeling especially towards persona-aware response generation**.

We introduce a novel persona-aware MPC response generation model which utilizes this dataset and provides a baseline towards this research area. **We find that modeling persona attributes as nodes and connecting them to utterance nodes via edges for modeling performs best, and produces persona-aware outputs, outperforming the non-persona-aware base model.** We aim to conduct further work with multimodal modeling to account for the images and gifs shared on Community Connect. We also aim to conduct experiments with multilingual data to study the effect of persona-aware generation in languages other than English.

We also study issues in NLG evaluation towards multi-party conversational AI. We find that there exist many inconsistencies in error reporting even when performance is reported on the same shared task. We also find that with the challenges introduced by multi-party conversation modeling, there is a need to introduce measures which can more effectively contribute towards gauging progress towards modeling. **We introduce an expanded taxonomy which introduces these new error categories, and utilize some of the guidelines towards evaluating our persona-aware model.** In future work, we aim to validate these error categories further, in keeping with previous work in this area with two-party dialogue model evaluation [2]. We also aim to collaborate with the community at large towards creating a benchmark which could help towards gauging the progress in the field in a more streamlined, robust manner.

REFERENCES

[1] S. Garg, B. Martinovski, S. Robinson, J. Stephan, J. Tetreault, and D. Traum, "Evaluation of transcription and annotation tools for a multi-modal, multi-party dialogue corpus," in *LREC*, 2004.

[2] R. Higashinaka, M. Araki, H. Tsukahara, and M. Mizukami, "Integrated taxonomy of errors in chat-oriented dialogue systems," in *SIGDIAL*, 2021.

[3] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 1, pp. 517–520, IEEE Computer Society, 1992.

[4] R. Rameshkumar and P. Bailey, "Storytelling with dialogue: A critical role dungeons and dragons dataset," in *ACL*, 2020.

[5] P. Sibun, "Beyond dialogue: the six ws of multi-party interaction," in *Working Notes of AAAI97 Spring Symposium On Mixed-Initiative Interaction, Stanford, CA*, pp. 145–150, 1997.

[6] K. Kirchhoff and M. Ostendorf, "Directions for multi-party human-computer interaction research," in *HLT-NAACL 2003*, 2003.

[7] D. Traum, "Issues in multiparty dialogues," in *Workshop on Agent Communication Languages*, 2003.

[8] H. Ouchi and Y. Tsuboi, "Addressee and response selection for multi-party conversation," in *EMNLP*, 2016.

[9] R. Zhang, H. Lee, L. Polymenakos, and D. R. Radev, "Addressee and response selection in multi-party conversations with speaker interaction rnns," in *AAAI*, 2018.

[10] R. Le, W. Hu, M. Shang, Z. You, L. Bing, D. Zhao, and R. Yan, "Who is speaking to whom? learning to identify utterance addressee in multi-party conversations," in *EMNLP/IJCNLP*, 2019.

[11] W. Wang, S. C. Hoi, and S. Joty, "Response selection for multi-party conversations with dynamic topic tracking," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6581–6591, 2020.

[12] J.-C. Gu, C. Tao, Z. Ling, C. Xu, X. Geng, and D. Jiang, "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 a nd the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3682–3692, 2021.

[13] F. Dignum and G. Vreeswijk, "Towards a testbed for multi-party dialogues," in *Workshop on Agent Communication Languages*, 2003.

[14] D. R. Traum, S. Robinson, and J. Stephan, "Evaluation of multi-party virtual reality dialogue interaction," in *LREC*, 2004.

[15] D. R. Traum, S. Robinson, and J. Stephan, "Evaluation of multi-party reality dialogue interaction," tech. rep., UNIVERSITY OF SOUTHERN CALIFORNIA MARINA DEL REY CA INST FOR CREATIVE TECHNOLOGIES, 2006.

[16] J.-C. Gu, C.-H. Tan, C. Tao, Z.-H. Ling, H. Hu, X. Geng, and D. Jiang, "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)*, pp. 5086–5097, 2022.

[17] S. Mallios and N. Bourbakis, "A survey on human machine dialogue systems," *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)*, pp. 1–7, 2016.

[18] H. Chen, X. Liu, D. Yin, and J. Tang, "A survey on dialogue systems: Recent advances and new frontiers," *Acm Sigkdd Explorations Newsletter*, vol. 19, no. 2, pp. 25–35, 2017.

[19] R. Perera and P. Nand, "Recent advances in natural language generation: A survey and classification of the empirical literature," *Computing and Informatics*, vol. 36, no. 1, pp. 1–32, 2017.

[20] A. Gatt and E. Krahmer, "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation," *Journal of Artificial Intelligence Research*, vol. 61, pp. 65–170, 2018.

[21] S. Santhanam and S. Shaikh, "A survey of natural language generation techniques with a focus on dialogue systems-past, present and future directions," *arXiv preprint arXiv:1906.00500*, 2019.

[22] Y. Ma, K. L. Nguyen, F. Z. Xing, and E. Cambria, "A survey on empathetic dialogue systems," *Information Fusion*, vol. 64, pp. 50–70, 2020.

[23] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, and X. Huang, "Pre-trained models for natural language processing: A survey," *Science China Technological Sciences*, pp. 1–26, 2020.

[24] C. Dong, Y. Li, H. Gong, M. Chen, J. Li, Y. Shen, and M. Yang, "A survey of natural language generation," *arXiv preprint arXiv:2112.11739*, 2021.

[25] J. Ni, T. Young, V. Pandelea, F. Xue, and E. Cambria, "Recent advances in deep learning based dialogue systems: A systematic survey," *Artificial intelligence review*, vol. 56, no. 4, pp. 3055–3155, 2023.

[26] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations. in naacl," 2018.

[27] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 328–339, 2018.

[28] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training."

[29] K. Elkins and J. Chun, "Can gpt-3 pass a writers turing test," *Journal of Cultural Analytics*, vol. 5, no. 2, p. 17212, 2020.

[30] R. OpenAI, "Gpt-4 technical report," *arXiv*, pp. 2303–08774, 2023.

[31] J. Li, M. Galley, C. Brockett, G. Spithourakis, J. Gao, and W. B. Dolan, "A persona-based neural conversation model," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 994–1003, 2016.

[32] B. Hedayatnia, S. Kim, Y. Liu, K. Gopalakrishnan, M. Eric, and D. Hakkani-Tur, "Policy-driven neural response generation for knowledge-grounded dialogue systems," in *INLG*, 2020.

[33] Q. Liu, Y. Chen, B. Chen, L. Jian-Guang, Z. Chen, B. Zhou, and D. Zhang, "You impress me: Dialogue generation via mutual persona perception," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1417–1427, 2020.

[34] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," *Advances in neural information processing systems*, vol. 27, 2014.

[35] S. Zarrieß, H. Voigt, and S. Schüz, "Decoding methods in neural language generation: A survey," *Information*, vol. 12, no. 9, p. 355, 2021.

[36] G. Caldarini, S. Jaf, and K. McGarry, "A literature survey of recent advances in chatbots," *Information*, vol. 13, no. 1, p. 41, 2022.

[37] R. Winkler, M. L. Neuweiler, E. Bittner, and M. Söllner, "Hey alexa, please help us solve this problem! how interactions with smart personal assistants improve group performance," in *ICIS International Conference of Information Systems*, (Munich), ACM Digital, 2019.

[38] I. Seeber, E. Bittner, R. O. Briggs, T. De Vreede, G.-J. De Vreede, A. Elkins, R. Maier, A. B. Merz, S. Oeste-Reiß, N. Randrup, *et al.*, "Machines as teammates: A research agenda on ai in team collaboration," *Information & management*, vol. 57, no. 2, p. 103174, 2020.

[39] A. Fioramonte and C. Vásquez, "Multi-party talk in the medical encounter: Socio-pragmatic functions of family members' contributions in the treatment advice phase," *Journal of Pragmatics*, vol. 139, pp. 132–145, 2019.

[40] S. Al Moubayed, J. Beskow, G. Skantze, and B. Granström, "Furhat: a back-projected human-like robot head for multiparty human-machine interaction," in *Cognitive behavioural systems*, pp. 114–130, Springer, 2012.

[41] G. Skantze, S. Moubayed, J. Gustafson, J. Beskow, and B. Granström, "Furhat at robotville : A robot head harvesting the thoughts of the public through multi-party dialogue," in *IVA 2012*, 2012.

[42] K. Mahajan, "A repository of citations related to multi-party conversations - past research and available resources," June 2023.

[43] Z. Zhang, R. Takanobu, Q. Zhu, M. Huang, and X. Zhu, "Recent advances and challenges in task-oriented dialog systems," *Science China Technological Sciences*, vol. 63, no. 10, pp. 2011–2027, 2020.

[44] Z. Zhang and H. Zhao, "Advances in multi-turn dialogue comprehension: A survey," *arXiv preprint arXiv:2103.03125*, 2021.

[45] M.-P. Huget and Y. Demazeau, "First steps towards multi-party communication," in *International Workshop on Agent Communication*, pp. 65–75, Springer, 2004.

[46] S. Robinson, B. Martinovski, S. Garg, J. Stephan, and D. Traum, "Issues in corpus development for multi-party multi-modal task-oriented dialogue," in *LREC*, 2004.

[47] F. Kronlid, "Steps towards multi-party dialogue management," 2008.

[48] H. P. Branigan, "Perspectives on multi-party dialogue," *Research on Language and Computation*, vol. 4, pp. 153–177, 2006.

[49] M. Porcheron, J. E. Fischer, and S. Sharples, ""do animals have accents?": Talking with agents in multi-party conversation," *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, 2017.

[50] G. Leech, "100 million words of english:the british national corpus (bnc)," *Second Language Research*, vol. 28, pp. 1–13, 1992.

[51] M. McCarthy, *Spoken language and applied linguistics*. Cambridge University Press, 1998.

[52] C. Oertel, F. Cummins, J. Edlund, P. Wagner, and N. Campbell, "D64: a corpus of richly recorded conversational interaction," *Journal on Multimodal User Interfaces*, vol. 7, pp. 19–28, 2012.

[53] A. Stupakov, E. Hanusa, D. Vijaywargi, D. Fox, and J. Bilmes, "The design and collection of cosine, a multi-microphone in situ speech corpus recorded in noisy environments," *Comput. Speech Lang.*, vol. 26, pp. 52–66, 2012.

[54] H. Hung and G. Chittaranjan, "The idiap wolf corpus: exploring group behaviour in a competitive role-playing game," *Proceedings of the 18th ACM international conference on Multimedia*, 2010.

[55] D. Litman, S. Paletz, Z. Rahimi, S. Allegretti, and C. Rice, "The teams corpus and entrainment in multi-party spoken dialogues," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, (Austin, Texas), pp. 1421–1431, Association for Computational Linguistics, 2016.

[56] A. Stenström and L. E. Breivik, "The bergen corpus of london teenager language (colt)," *ICAME journal*, vol. 17, p. 128, 1993.

[57] R. C. Simpson-Vlach and S. Leicher, *The MICASE handbook: A resource for users of the Michigan corpus of academic spoken English*. University of Michigan Press ELT, 2006.

[58] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the ami and amida projects," *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pp. 238–247, 2007.

[59] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The icsi meeting corpus," *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03).*, vol. 1, pp. I–I, 2003.

[60] E. Shriberg, R. Dhillon, S. Bhagat, J. Ang, and H. Carvey, "The ICSI meeting recorder dialog act (MRDA) corpus," in *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, (Cambridge, Massachusetts, USA), pp. 97–100, Association for Computational Linguistics, 2004.

[61] Z. Yang, B. Li, Y. Zhu, I. King, G.-A. Levow, and H. Meng, "Collection of user judgments on spoken dialog system with crowdsourcing," *2010 IEEE Spoken Language Technology Workshop*, pp. 277–282, 2010.

[62] M. Barlow, *Corpus of Spoken, Professional American-English*. Rice University, 2000.

[63] M. Kytö and T. Walker, *Guide to A corpus of English dialogues 1560-1760*. Acta Universitatis Upsaliensis, 2006.

[64] C. Zhu, Y. Liu, J. Mei, and M. Zeng, "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*, (Online), pp. 5927–5934, Association for Computational Linguistics, 2021.

[65] B. P. Majumder, S. Li, J. Ni, and J. McAuley, "Interview: A large-scale open-source corpus of media dialog," *arXiv preprint arXiv:2004.03090*, 2020.

[66] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin, "Canal9: A database of political debates for analysis of social interactions," *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pp. 1–4, 2009.

[67] R. E. Banchs, "Movie-DiC: a movie dialogue corpus for research and development," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, (Jeju Island, Korea), pp. 203–207, Association for Computational Linguistics, 2012.

[68] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," in *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, (Portland, Oregon, USA), pp. 76–87, Association for Computational Linguistics, 2011.

[69] "Film scripts online series." Accessed on June 29, 2023.

[70] J. Tiedemann, "Parallel data, tools and interfaces in OPUS," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, (Istanbul, Turkey), pp. 2214–2218, European Language Resources Association (ELRA), 2012.

[71] D. Ameixa and L. Coheur, "From subtitles to human interactions : introducing the subtle corpus," tech. rep., Technical report, 2013.

[72] M. Walker, G. Lin, and J. Sawyer, "An annotated corpus of film dialogue for learning and characterizing character style," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, (Istanbul, Turkey), pp. 1373–1378, European Language Resources Association (ELRA), 2012.

[73] M. Davies, *Corpus of american soap operas*. Brigham Young University, 2013.

[74] A. Roy, C. Guinaudeau, H. Bredin, and C. Barras, "TVD: A reproducible and multiply aligned TV series dataset," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, (Reykjavik, Iceland), pp. 418–425, European Language Resources Association (ELRA), 2014.

[75] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 527–536, Association for Computational Linguistics, 2019.

[76] X. Bost, V. Labatut, and G. Linares, "Serial speakers: a dataset of TV series," in *Proceedings of the 12th Language Resources and Evaluation Conference*, (Marseille, France), pp. 4256–4264, European Language Resources Association, 2020.

[77] M. Firdaus, H. Chauhan, A. Ekbal, and P. Bhattacharyya, "MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations," in *Proceedings of the 28th International Conference on Computational Linguistics*, (Barcelona, Spain (Online)), pp. 4441–4453, International Committee on Computational Linguistics, 2020.

[78] E. N. Forsythand and C. H. Martell, "Lexical and discourse analysis of online chat dialog," *International Conference on Semantic Computing (ICSC 2007)*, pp. 19–26, 2007.

[79] R. Lowe, N. Pow, I. V. Serban, and J. Pineau, "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*, pp. 285–294, 2015.

[80] D. C. Uthus and D. Aha, "The ubuntu chat corpus for multiparticipant chat analysis," in *AAAI Spring Symposium: Analyzing Microtext*, 2013.

[81] J. Li, M. Liu, M.-Y. Kan, Z. Zheng, Z. Wang, W. Lei, T. Liu, and B. Qin, "Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure," in *Proceedings of the 28th International Conference on Computational Linguistics*, (Barcelona, Spain (Online)), pp. 2642–2652, International Committee on Computational Linguistics, 2020.

[82] S. Shaikh, T. Strzalkowski, A. Broadwell, J. Stromer-Galley, S. Taylor, and N. Webb, "MPC: A multi-party chat corpus for modeling social phenomena in discourse," in *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, (Valletta, Malta), European Language Resources Association (ELRA), 2010.

[83] S. Afantenos, N. Asher, F. Benamara, A. Cadilhac, C. Dégremont, P. Denis, M. Guhe, S. Keizer, A. Lascarides, O. Lemon, *et al.*, "Developing a corpus of strategic conversation in the settlers of catan," in *SeineDial 2012-The 16th Workshop On The Semantics and Pragmatics Of Dialogue*, 2012.

[84] A. Djalali, S. Lauer, and C. Potts, "Corpus evidence for preference-driven interpretation," in *Amsterdam Colloquium on Logic, Language and Meaning*, 2011.

[85] J. P. Chang, C. Chiam, L. Fu, A. Wang, J. Zhang, and C. Danescu-Niculescu-Mizil, "ConvoKit: A toolkit for the analysis of conversations," in *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, (1st virtual meeting), pp. 57–60, Association for Computational Linguistics, 2020.

[86] N. Schrading, C. Ovesdotter Alm, R. Ptucha, and C. Homan, "An analysis of domestic abuse discourse on Reddit," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (Lisbon, Portugal), pp. 2577–2583, Association for Computational Linguistics, 2015.

[87] M. Walker, J. F. Tree, P. Anand, R. Abbott, and J. King, "A corpus for research on deliberation and debate," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, (Istanbul, Turkey), pp. 812–817, European Language Resources Association (ELRA), 2012.

[88] J. Andreas, S. Rosenthal, and K. McKeown, "Annotating agreement and disagreement in threaded discussion," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, (Istanbul, Turkey), pp. 818–822, European Language Resources Association (ELRA), 2012.

[89] S. Rosenthal and K. McKeown, "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*, (Prague, Czech Republic), pp. 168–177, Association for Computational Linguistics, 2015.

[90] A. Ritter, C. Cherry, and B. Dolan, "Unsupervised modeling of Twitter conversations," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, (Los Angeles, California), pp. 172–180, Association for Computational Linguistics, 2010.

[91] C. Shaoul and C. Westbury, "A usenet corpus (2005–2007)," *University of Alberta, Edmonton, AB*, 2007.

[92] C. Shaoul and C. Westbury, "A usenet corpus (2005-2010)," *Edmonton, AB: University of Alberta*, 2011.

[93] G. Leech, R. Garside, and M. Bryant, "CLAWS4: The tagging of the British National Corpus," in *COLING 1994 Volume 1: The 15th International Conference on Computational Linguistics*, 1994.

[94] P. Rayson, G. Leech, and M. Hodges, "Social differentiation in the use of english vocabulary: some analyses of the conversational component of the british national corpus," *International Journal of Corpus Linguistics*, vol. 2, pp. 133–152, 1997.

[95] G. Leech, P. Rayson, and A. Wilson, *Word frequencies in written and spoken English: based on the British National Corpus.* Longman, 2001.

[96] R. Xiao and H. Tao, "A corpus-based sociolinguistic study of amplifiers in british english," *Sociolinguistic Studies*, vol. 1, pp. 241–273, 2007.

[97] A. O'keeffe, M. McCarthy, and R. Carter, *From corpus to classroom: Language use and language teaching*. Cambridge University Press, 2007.

[98] A.-B. Stenström, G. Andersen, and I. K. Hasund, *Trends in Teenage Talk: Corpus compilation, analysis and findings*, vol. 8. John Benjamins Publishing, 2002.

[99] C. Oertel, G. Castellano, M. Chetouani, J. Nasir, M. Obaid, C. Pelachaud, and C. Peters, "Engagement in human-agent interaction: An overview," *Frontiers in Robotics and AI*, vol. 7, p. 92, 2020.

[100] P. A. Raffensperger, R. Webb, P. Bones, and A. McInnes, "A simple metric for turn-taking in emergent communication," *Adaptive Behavior*, vol. 20, pp. 104 – 116, 2012.

[101] Y. Kano, C. Aranha, M. Inaba, F. Toriumi, H. Osawa, D. Katagami, T. Otsuki, I. Tsunoda, S. Nagayama, D. Tellols, Y. Sugawara, and Y. Nakata, "Overview of AIWolfDial 2019 shared task: Contest of automatic dialog agents to play the werewolf game through conversations," in *Proceedings of the 1st International Workshop of AI Werewolf and Dialog System (AIWolfDial2019)*, (Tokyo, Japan), pp. 1–6, Association for Computational Linguistics, 2019.

[102] Z. Rahimi and D. Litman, "Entrainment2vec: Embedding entrainment for multi-party dialogues," in *AAAI*, 2020.

[103] D. Liu, "The most frequently used spoken american english idioms: A corpus analysis and its implications.," *TESOL Quarterly*, vol. 37, pp. 671–700, 2003.

[104] G. Murray and S. Renals, "Towards online speech summarization," in *INTERSPEECH*, 2007.

[105] "Npr." Accessed on June 29, 2023.

[106] "Cnn." Accessed on June 29, 2023.

[107] J. Zhang, R. Kumar, S. Ravi, and C. Danescu-Niculescu-Mizil, "Conversational flow in Oxford-style debates," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (San Diego, California), pp. 136–141, Association for Computational Linguistics, 2016.

[108] S. Kim, F. Valente, and A. Vinciarelli, "Automatic detection of conflicts in spoken conversations: Ratings and analysis of broadcast political debates," *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5089–5092, 2012.

[109] U. Lenker, "'there's an issue there...': Signalling functions of discourse-deictic there in the history of english," *Language Sciences*, vol. 68, pp. 94–105, 2018.

[110] G. Ji and J. Bilmes, "Multi-speaker language modeling," in *Proceedings of HLT-NAACL 2004: Short Papers*, (Boston, Massachusetts, USA), pp. 133–136, Association for Computational Linguistics, 2004.

[111] C. Zhu, R. Xu, M. Zeng, and X. Huang, "End-to-end abstractive summarization for meetings," *ArXiv*, vol. abs/2004.02016, 2020.

[112] J. Demmen, *A corpus stylistic investigation of the language style of Shakespeare's plays in the context of other contemporaneous plays*. Lancaster University, 2012.

[113] L. Wang, X. Zhang, Z. Tu, A. Way, and Q. Liu, "Automatic construction of discourse corpora for dialogue translation," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, (Portorož, Slovenia), pp. 2748–2754, European Language Resources Association (ELRA), 2016.

[114] N. Asghar, I. Kobyzev, J. Hoey, P. Poupart, and M. B. Sheikh, "Generating emotionally aligned responses in dialogues using affect control theory," *arXiv preprint arXiv:2003.03645*, 2020.

[115] "Imsdb." Accessed on June 29, 2023.

[116] W. Xu, C. Hargood, W. Tang, and F. Charles, "Towards generating stylistic dialogues for narratives using data-driven approaches," in *ICIDS*, 2018.

[117] "Opensubtitles." Accessed on June 29, 2023.

[118] P. Lison and J. Tiedemann, "OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, (Portorož, Slovenia), pp. 923–929, European Language Resources Association (ELRA), 2016.

[119] P. Lison, J. Tiedemann, and M. Kouylekov, "OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, (Miyazaki, Japan), European Language Resources Association (ELRA), 2018.

[120] M. S. Khaghaninejad, M. Dehbozorgi, and M. A. Mokhtari, "Cultural representations of americans, europeans, africans and arabs in american soap operas: A corpus-based analysis," *Language & Translation*, vol. 7, no. 3, pp. 133–141, 2019.

[121] E. Knyazeva, G. Wisniewski, H. Bredin, and F. Yvon, "Structured prediction for speaker identification in tv series," in *INTERSPEECH*, 2015.

[122] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "EmotionLines: An emotion corpus of multi-party conversations," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, (Miyazaki, Japan), European Language Resources Association (ELRA), 2018.

[123] T. Saha, A. Patra, S. Saha, and P. Bhattacharyya, "Towards emotion-aided multi-modal dialogue act classification," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Online), pp. 4361–4372, Association for Computational Linguistics, 2020.

[124] H. Jiang, X. Zhang, and J. D. Choi, "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*, no. Volume 34, Number 10, pp. 13821–13822, 2020.

[125] H. Christian, D. Suhartono, A. Chowanda, and K. Z. Zamli, "Text based personality prediction from multiple social media data sources using pre-trained language model and model averaging," *Journal of Big Data*, vol. 8, no. 1, pp. 1–20, 2021.

[126] F. Yang, X. Quan, Y. Yang, and J. Yu, "Multi-document transformer for personality detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, no. Volume 35, Number 16, pp. 14221–14229, 2021.

[127] E. Loper and S. Bird, "NLTK: The natural language toolkit," in *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, (Philadelphia, Pennsylvania, USA), pp. 63–70, Association for Computational Linguistics, 2002.

[128] T. Strzalkowski, S. Shaikh, T. Liu, G. Broadwell, J. Stromer-Galley, S. Taylor, U. Boz, V. Ravishankar, and X. Ren, "Modeling leadership and influence in multi-party online discourse," in *COLING*, 2012.

[129] "Ubuntu irc chatroom." Accessed on June 29, 2023.

[130] R. Lowe, N. Pow, I. Serban, L. Charlin, C. Liu, and J. Pineau, "Training end-to-end dialogue systems with the ubuntu dialogue corpus," *Dialogue Discourse*, vol. 8, pp. 31–65, 2017.

[131] "Reddit." Accessed on June 29, 2023.

[132] "Wikipedia." Accessed on June 29, 2023.

[133] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," in *ICWSM*, 2020.

[134] A. Cadilhac, N. Asher, F. Benamara, and A. Lascarides, "Grounding strategic conversation: Using negotiation dialogues to predict trades in a win-lose game," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (Seattle, Washington, USA), pp. 357–368, Association for Computational Linguistics, 2013.

[135] "Createdebate." Accessed on June 29, 2023.

[136] G. Rakshit, K. K. Bowden, L. Reed, A. Misra, and M. Walker, "Debbie, the debate bot of the future," in *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems*, vol. 510, p. 45, Springer, 2018.

[137] "Usenet." Accessed on June 29, 2023.

[138] "Twitter." Accessed on June 29, 2023.

[139] D. Traum, S. Marsella, J. Gratch, J. Lee, and A. Hartholt, "Multi-party, multi-issue, multi-strategy negotiation for multi-modal virtual agents," in *IVA*, 2008.

[140] J. Lee, S. Marsella, D. Traum, J. Gratch, and B. Lance, "The rickel gaze model: A window on the mind of a virtual human," in *International workshop on intelligent virtual agents*, pp. 296–303, Springer, 2007.

[141] N. Jovanovic, R. op den Akker, and A. Nijholt, "A corpus for studying addressing behaviour in multi-party dialogues," *Language Resources and Evaluation*, vol. 40, no. 1, pp. 5–23, 2006.

[142] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia, *et al.*, "The chil audiovisual corpus for lecture and meeting analysis inside smart rooms," *Language resources and evaluation*, vol. 41, no. 3, pp. 389–407, 2007.

[143] C.-S. Wu, A. Madotto, W. Liu, P. Fung, and C. Xiong, "Qaconv: Question answering on informative conversations," *arXiv preprint arXiv:2105.06912*, 2021.

[144] L. Chen, R. Rose, Y. Qiao, I. Kimbara, F. Parrill, H. Welji, T. X. Han, J. Tu, Z. Huang, M. Harper, F. K. H. Quek, Y. Xiong, D. McNeill, R. Tuttle, and T. Huang, "Vace multimodal meeting corpus," in *MLMI*, 2005.

[145] M. Koutsombogera and C. Vogel, "Modeling collaborative multimodal behavior in group dialogues: The MULTISIMO corpus," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, (Miyazaki, Japan), European Language Resources Association (ELRA), 2018.

[146] N. Mana, B. Lepri, P. Chippendale, A. Cappelletti, F. Pianesi, P. Svaizer, and M. Zancanaro, "Multimodal corpus of multi-party meetings for automatic social behavior analysis and personality traits detection," *The Medical Roundtable*, pp. 9–14, 2007.

[147] P. Shah, D. Hakkani-Tür, G. Tür, A. Rastogi, A. Bapna, N. Nayak, and L. Heck, "Building a conversational agent overnight with dialogue self-play," *arXiv preprint arXiv:1801.04871*, 2018.

[148] "Chatarena." Accessed on June 30, 2023.

[149] D. McCarthy, B. Keller, and R. Navigli, "Getting synonym candidates from raw data in the english lexical substitution task," in *Proceedings of the 14th euralex international congress*, 2010.

[150] W. Roberts and E. Erklärung, *Integrating syntax and semantics for word sense disambiguation*. PhD thesis, MA thesis. Saarbrücken, Germany: Universität des Saarlandes, 2011.

[151] J. D. Choi and H. Y. Chen, "SemEval 2018 task 4: Character identification on multiparty dialogues," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, (New Orleans, Louisiana), pp. 57–64, Association for Computational Linguistics, 2018.

[152] "Truman platform." Accessed on June 29, 2023.

[153] K. Mahajan, S. R. Choudhury, S. M. Levens, T. Gallicano, and S. Shaikh, "Community connect: A mock social media platform to study online behavior," *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, 2021.

[154] T. Ricento, "Clausal ellipsis in multi-party conversation in english," *Journal of Pragmatics*, vol. 11, pp. 751–775, 1987.

[155] M. Ishizaki and T. Kato, "Exploring the characteristics of multi-party dialogues," in *COLING-ACL*, 1998.

[156] S. Yoon and S. Brown-Schmidt, "Contextual integration in multiparty audience design," *Cognitive science*, vol. 43 12, p. e12807, 2019.

[157] P. M. Aoki, M. H. Szymanski, L. Plurkowski, J. D. Thornton, A. Woodruff, and W. Yi, "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*, pp. 393–402, 2006.

[158] C. Howes, M. Purver, P. Healey, G. Mills, and E. Gregoromichelaki, "On incrementality in dialogue: Evidence from compound contributions," *Dialogue Discourse*, vol. 2, pp. 279–311, 2011.

[159] P. McBurney, D. Hitchcock, and S. Parsons, "The eightfold way of deliberation dialogue," *International Journal of Intelligent Systems*, vol. 22, 2007.

[160] D. Hitchcock, "Some principles of rational mutual inquiry," in *On Reasoning and Argument*, pp. 313–321, Springer, 2017.

[161] P. G. Healey and G. Mills, "Participation, precedence and co-ordination in dialogue," in *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, vol. 320, Citeseer, 2006.

[162] S. E. Brennan and H. H. Clark, "Conceptual pacts and lexical choice in conversation.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 22, no. 6, p. 1482, 1996.

[163] M. J. Pickering and S. Garrod, "The interactive-alignment model: Developments and refinements," *Behavioral and Brain Sciences*, vol. 27, no. 2, p. 212, 2004.

[164] A. Eshghi and P. G. Healey, "What is conversation? distinguishing dialogue contexts," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 31, 2009.

[165] A. Kendon, *Conducting interaction: Patterns of behavior in focused encounters*, vol. 7. CUP Archive, 1990.

[166] T. Strzalkowski, G. Broadwell, J. Stromer-Galley, S. Shaikh, S. Taylor, and N. Webb, "Modeling socio-cultural phenomena in discourse," in *COLING*, 2010.

[167] N. Jovanovic and R. Akker, "Towards automatic addressee identification in multi-party dialogues," in *SIGDIAL Workshop*, 2004.

[168] T. Hawes, J. Lin, and P. Resnik, "Elements of a computational model for multiparty discourse: The turn-taking behavior of supreme court justices," *J. Assoc. Inf. Sci. Technol.*, vol. 60, pp. 1607–1615, 2009.

[169] K. Ma, C. Xiao, and J. D. Choi, "Text-based speaker identification on multiparty dialogues using multi-document convolutional neural networks," in *ACL*, 2017.

[170] M. G. de Bayser, P. Cavalin, C. Pinhanez, and B. Zadrozny, "Learning multiparty turn-taking models from dialogue logs," *ArXiv*, vol. abs/1907.02090, 2019.

[171] M. G. de Bayser, M. A. Guerra, P. Cavalin, and C. Pinhanez, "A hybrid solution to learn turn-taking in multi-party service-based chat groups," *Interactions*, vol. 10, no. 4, p. 2, 2020.

[172] L. Qiu, Y. Zhao, W. Shi, Y. Liang, F. Shi, T. Yuan, Z. Yu, and S.-c. Zhu, "Structured attention for unsupervised dialogue structure induction," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1889–1899, 2020.

[173] J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2980–2988, 2015.

[174] W. Hu, Z. Chan, B. Liu, D. Zhao, J. Ma, and R. Yan, "Gsn: A graph-structured network for multi-party dialogues," in *IJCAI*, 2019.

[175] R. Zhang, L. Polymenakos, D. Radev, D. Nahamoo, and L. Honglak, "Modeling multiparty conversation dynamics: speaker, response, addressee selection using a novel deep learning approach," 2020. US Patent 10,657,962.

[176] Q. Jia, Y. Liu, S. Ren, K. Zhu, and H. Tang, "Multi-turn response selection using dialogue dependency relations," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1911–1920, 2020.

[177] K. Mahajan, S. Santhanam, and S. Shaikh, "Towards evaluation of multi-party dialogue systems," in *Proceedings of the 15th International Conference on Natural Language Generation*, (Waterville, Maine, USA and virtual meeting), pp. 278–287, Association for Computational Linguistics, July 2022.

[178] J.-C. Gu, T. Li, Q. Liu, Z.-H. Ling, Z. Su, S. Wei, and X. Zhu, "Speaker-aware bert for multi-turn response selection in retrieval-based chatbots," in *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pp. 2041–2044, 2020.

[179] H. Zhang, Z. Chan, Y. Song, D. Zhao, and R. Yan, "When less is more: Using less context information to generate better utterances in group conversations," in *NLPCC*, 2018.

[180] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *ArXiv*, vol. abs/1412.3555, 2014.

[181] C. Liu, K. Liu, S. He, Z. Nie, and J. Zhao, "Incorporating interlocutor-aware context into response generation on multi-party chatbots," in *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pp. 718–727, 2019.

[182] K. Mahajan and S. Shaikh, "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*, pp. 338–352, 2021.

[183] "Praat." Accessed on June 29, 2023.

[184] "Anvil." Accessed on June 29, 2023.

[185] "Transcriber." Accessed on June 29, 2023.

[186] J. Besser and J. Alexandersson, "A comprehensive disfluency model for multi-party interaction," in *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pp. 182–189, 2007.

[187] V. Popescu and J. Caelen, "A rhetorical structuring model for natural language generation in human-computer multi-party dialogue," 2008.

[188] A. Lascarides and N. Asher, "Segmented discourse representation theory: Dynamic semantics with discourse structure," in *Computing meaning*, pp. 87–124, Springer, 2008.

[189] J. Li, M. Liu, B. Qin, Z. Zheng, and T. Liu, "An annotation scheme of a large-scale multi-party dialogues dataset for discourse parsing and machine comprehension," *ArXiv*, vol. abs/1911.03514, 2019.

[190] E. Gilmartin and N. Campbell, "Capturing chat: Annotation and tools for multiparty casual conversation," in *LREC*, 2016.

[191] E. Gilmartin, C. Vogel, and N. Campbell, "Chats and chunks: Annotation and analysis of multiparty long casual conversations," in *LREC*, 2018.

[192] V.-A. Nguyen, Y. Hu, J. L. Boyd-Graber, and P. Resnik, "Argviz: Interactive visualization of topic dynamics in multi-party conversations," in *HLT-NAACL*, 2013.

[193] V.-A. Nguyen, J. L. Boyd-Graber, and P. Resnik, "Sits: A hierarchical nonparametric model using speaker identity for topic segmentation in multiparty conversations," in *ACL*, 2012.

[194] S. Sheehan, P. Albert, S. Luz, and M. Masoodian, "Temoco: A visualization tool for temporal analysis of multi-party dialogues in clinical settings," *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 690–695, 2019.

[195] A. Popescu-Belis, "Dialogue acts: One or more dimensions," *ISSCO Working-Paper*, vol. 62, pp. 1–46, 2005.

[196] J. Allen and M. Core, "Draft of damsl: Dialog act markup in several layers," 1997.

[197] D. Jurafsky, "Switchboard swbd-damsl shallow-discourse-function annotation coders manual," *Institute of Cognitive Science Technical Report*, 1997.

[198] A. Clark and A. Popescu-Belis, "Multi-level dialogue act tags," in *SIGDIAL Workshop*, 2004.

[199] N. Jovanovic, R. Akker, and A. Nijholt, "A corpus for studying addressing behaviour in multi-party dialogues," *Language Resources and Evaluation*, vol. 40, pp. 5–23, 2005.

[200] H. Bunt, "The dit++ taxonomy for functional dialogue markup," in *AAMAS 2009 Workshop, Towards a Standard Markup Language for Embodied Dialogue Acts*, pp. 13–24, 2009.

[201] H. Bunt, J. Alexandersson, J.-W. Choe, A. C. Fang, K. Hasida, V. Petukhova, A. Popescu-Belis, and D. Traum, "Iso 24617-2: A semantically-based standard for dialogue annotation," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 430–437, 2012.

[202] "Ditplusplus." Accessed on June 29, 2023.

[203] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, p. 100, Citeseer, 2005.

[204] D. Traum, C. Henry, S. M. Lukin, R. Artstein, F. Gervits, K. A. Pollard, C. Bonial, S. Lei, C. R. Voss, M. Marge, C. J. Hayes, and S. G. Hill, "Dialogue structure annotation for multi-floor interaction," in *LREC*, 2018.

[205] S. Kim, L. Cavedon, and T. Baldwin, "Classifying dialogue acts in multi-party live chats," in *PACLIC*, 2012.

[206] M. Tavafi, Y. Mehdad, S. R. Joty, G. Carenini, and R. Ng, "Dialogue act recognition in synchronous and asynchronous conversations," in *SIGDIAL Conference*, 2013.

[207] D. Amanova, V. Petukhova, and D. Klakow, "Creating annotated dialogue resources: Cross-domain dialogue act classification," in *LREC*, 2016.

[208] O. Irsoy, R. Gosangi, H. Zhang, W. Mu-Hsin, P. J. Lund, D. Pappadopulo, B. M. Fahy, N. Nephytou, and C. O. Diaz, "Dialogue act classification in group chats with dag-lstms," Oct. 28 2021. US Patent App. 17/234,745.

[209] G. Lerner, "On the place of linguistic resources in the organization of talk-in-interaction: 'second person' reference in multi-party conversation," *Pragmatics*, vol. 6, pp. 281–294, 1996.

[210] S. Gupta, J. Niekrasz, M. Purver, and D. Jurafsky, "Resolving "you" in multi-party dialog," in *SIGdial*, 2007.

[211] J. D. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proceedings of the Eighteenth International Conference on Machine Learning*, pp. 282–289, 2001.

[212] M. Frampton, R. Fernández, P. Ehlen, C. M. Christoudias, T. Darrell, and S. Peters, "Who is "you"? combining linguistic and gaze features to resolve second-person references in dialogue," in *EACL*, 2009.

[213] C. Müller, "Automatic detection of nonreferential it in spoken multi-party dialog," in *EACL*, 2006.

[214] W. W. Cohen, "Fast effective rule induction," in *Machine learning proceedings 1995*, pp. 115–123, Elsevier, 1995.

[215] C. Muller, "Resolving it, this, and that in unrestricted multi-party dialog," in *ACL 2007*, 2007.

[216] "Coreference resolution sota." Accessed on June 29, 2023.

[217] H. Branigan, M. Pickering, J. Mclean, and A. Cleland, "Syntactic alignment and participant role in dialogue," *Cognition*, vol. 104, pp. 163–197, 2007.

[218] H. Friedberg, D. Litman, and S. B. F. Paletz, "Lexical entrainment and success in student engineering groups," *2012 IEEE Spoken Language Technology Workshop (SLT)*, pp. 404–409, 2012.

[219] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of ACL-08: HLT, Short Papers*, pp. 169–172, 2008.

[220] Z. Rahimi, *Linguistic Entrainment in Multi-Party Spoken Dialogues*. PhD thesis, University of Pittsburgh, 2019.

[221] M. Yu, D. Litman, and S. Paletz, "Investigating the relationship between multi-party linguistic entrainment, team characteristics and the perception of team social outcomes," in *FLAIRS Conference*, 2019.

[222] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[223] C. Danescu-Niculescu-Mizil, L. Lee, B. Pang, and J. Kleinberg, "Echoes of power: Language effects and power differences in social interaction," in *Proceedings of the 21st international conference on World Wide Web*, pp. 699–708, 2012.

[224] S. Wasserman, K. Faust, C. U. Press, M. Granovetter, U. of Cambridge, and D. Iacobucci, *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

[225] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web.," tech. rep., Stanford InfoLab, 1999.

[226] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *NIPS*, 2013.

[227] S. Kim and T. Baldwin, "Extracting keywords from multi-party live chats," in *PACLIC*, 2012.

[228] M. Purver, K. P. Körding, T. Griffiths, and J. Tenenbaum, "Unsupervised topic modelling for multi-party spoken discourse," in *ACL*, 2006.

[229] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[230] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

[231] M. Georgescul, A. Clark, and S. Armstrong, "A comparative study of mixture models for automatic topic segmentation of multiparty dialogues," in *IJCNLP*, 2008.

[232] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine learning*, vol. 42, no. 1, pp. 177–196, 2001.

[233] R. Nishimura, Y. Todo, K. Yamamoto, and S. Nakagawa, "Chat-like spoken dialog system for a multi-party dialog incorporating two agents and a user," *Proceedings of iHAI*, vol. 13, 2013.

[234] A. Kai and S. Nakagawa, "A frame-synchronous continuous speech recognition algorithm using a top-down parsing of context-free grammar," in *Second International Conference on Spoken Language Processing*, 1992.

[235] R. Nishimura and S. Nakagawa, "Response timing generation and response type selection for a spontaneous spoken dialog system," in *2009 IEEE Workshop on Automatic Speech Recognition & Understanding*, pp. 462–467, IEEE, 2009.

[236] V. R. Martinez and J. Kennedy, "A multiparty chat-based dialogue system with concurrent conversation tracking and memory," *Proceedings of the 2nd Conference on Conversational User Interfaces*, 2020.

[237] G. Tür, A. Stolcke, L. L. Voss, J. Dowding, B. Favre, R. Fernández, M. Frampton, M. W. Frandsen, C. Frederickson, M. Graciarena, D. Z. Hakkani-Tür, D. Kintzing, K. Leveque, S. Mason, J. Niekrasz, S. Peters, M. Purver, K. Riedhammer, E. Shriberg, J. Tien, D. Vergyri, and F. Yang, "The calo meeting speech recognition and understanding system," *2008 IEEE Spoken Language Technology Workshop*, pp. 69–72, 2008.

[238] G. Tür, A. Stolcke, L. L. Voss, S. Peters, D. Z. Hakkani-Tür, J. Dowding, B. Favre, R. Fernández, M. Frampton, M. W. Frandsen, C. Frederickson, M. Graciarena, D. Kintzing, K. Leveque, S. Mason, J. Niekrasz, M. Purver, K. Riedhammer, E. Shriberg, J. Tien, D. Vergyri, and F. Yang, "The calo meeting assistant system," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, pp. 1601–1611, 2010.

[239] A. Stolcke, X. Anguera, K. Boakye, Ö. Çetin, F. Grézl, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: The icsi-sri spring 2005 speech-to-text evaluation system," in *International Workshop on Machine Learning for Multimodal Interaction*, pp. 463–475, Springer, 2005.

[240] M. Purver, J. Dowding, J. Niekrasz, P. Ehlen, S. Noorbaloochi, and S. Peters, "Detecting and summarizing action items in multi-party dialogue," in *SIGdial*, 2007.

[241] R. Fernández, M. Frampton, P. Ehlen, M. Purver, and S. Peters, "Modelling and detecting decisions in multi-party dialogue," in *SIGDIAL Workshop*, 2008.

[242] K. Riedhammer, B. Favre, and D. Hakkani-Tur, "A keyphrase based approach to interactive meeting summarization," in *2008 IEEE Spoken Language Technology Workshop*, pp. 153–156, IEEE, 2008.

[243] M. G. de Bayser, P. Cavalin, R. Souza, A. Braz, H. Candello, C. Pinhanez, and J. Briot, "A hybrid architecture for multi-party conversational systems," *ArXiv*, vol. abs/1705.01214, 2017.

[244] "Akka." Accessed on June 29, 2023.

[245] Q. Yang, Z. He, Z. Zhan, J. Zhao, Y. Zhang, and C. Hu, "Mids: End-to-end personalized response generation in untrimmed multi-role dialogue*," *2019 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, 2019.

[246] "Wechat." Accessed on June 30, 2023.

[247] Q. Zhu, Z. Zhang, Y. Fang, X. Li, R. Takanobu, J. chao Li, B. Peng, J. Gao, X. Zhu, and M. Huang, "Convlab-2: An open-source toolkit for building, evaluating, and diagnosing dialogue systems," in *ACL*, 2020.

[248] S. Lee, Q. Zhu, R. Takanobu, Z. Zhang, Y. Zhang, X. Li, J. Li, B. Peng, X. Li, M. Huang, *et al.*, "Convlab: Multi-domain end-to-end dialog system platform," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 64–69, 2019.

[249] F. Mairesse, M. Gasic, F. Jurcicek, S. Keizer, B. Thomson, K. Yu, and S. Young, "Spoken language understanding from unaligned data using discriminative classification models," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4749–4752, IEEE, 2009.

[250] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, pp. 4171–4186, 2019.

[251] O. Ramadan, P. Budzianowski, and M. Gasic, "Large-scale multi-domain belief tracking with knowledge sharing," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 432–437, 2018.

[252] H. Lee, J. Lee, and T.-Y. Kim, "Sumbt: Slot-utterance matching for universal and scalable belief tracking," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5478–5483, 2019.

[253] C.-S. Wu, A. Madotto, E. Hosseini-Asl, C. Xiong, R. Socher, and P. Fung, "Transferable multi-domain state generator for task-oriented dialogue systems," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 808–819, 2019.

[254] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[255] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[256] R. Takanobu, H. Zhu, and M. Huang, "Guided dialog policy learning: Reward estimation for multi-domain task-oriented dialog," 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)*, pp. 100–110, 2019.

[257] T.-H. Wen, M. Gasic, N. Mrkšić, P.-H. Su, D. Vandyke, and S. Young, "Semantically conditioned lstm-based natural language generation for spoken dialogue systems," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1711–1721, 2015.

[258] P. Budzianowski, I. Casanueva, B.-H. Tseng, and M. Gasic, "Towards end-to-end multi-domain dialogue modelling," 2018.

[259] W. Chen, J. Chen, P. Qin, X. Yan, and W. Y. Wang, "Semantically conditioned dialog response generation via hierarchical disentangled self-attention," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3696–3709, 2019.

[260] T. Zhao, K. Xie, and M. Eskenazi, "Rethinking action spaces for reinforcement learning in end-to-end dialog agents with latent variable models," 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)*, pp. 1208–1218, 2019.

[261] J. Schatzmann, B. Thomson, K. Weilhammer, H. Ye, and S. Young, "Agenda-based user simulation for bootstrapping a pomdp dialogue system," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers*, pp. 149–152, 2007.

[262] I. Gür, D. Hakkani-Tür, G. Tür, and P. Shah, "User modeling for task oriented dialogues," in *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 900–906, IEEE, 2018.

[263] W. Lei, X. Jin, M.-Y. Kan, Z. Ren, X. He, and D. Yin, "Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1437–1447, 2018.

[264] Y. Zhang, Z. Ou, and Z. Yu, "Task-oriented dialog systems that consider multiple appropriate responses under the same context," in *AAAI*, 2020.

[265] M. Lewis, D. Yarats, Y. Dauphin, D. Parikh, and D. Batra, "Deal or no deal? end-to-end learning of negotiation dialogues," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2443–2453, 2017.

[266] Q. Zhu, C. Geishauser, H.-c. Lin, C. van Niekerk, B. Peng, Z. Zhang, M. Heck, N. Lubis, D. Wan, X. Zhu, *et al.*, "Convlab-3: A flexible dialogue system toolkit based on a unified data format," *arXiv preprint arXiv:2211.17148*, 2022.

[267] A. Papangelis, M. Namazifar, C. Khatri, Y.-C. Wang, P. Molino, and G. Tur, "Plato dialogue system: A flexible conversational ai research platform," *ArXiv*, vol. abs/2001.06463, 2020.

[268] B. Morgan, F. Keshtkar, Y. Duan, P. Nash, and A. Graesser, "Using state transition networks to analyze multi-party conversations in a serious game," in *ITS*, 2012.

[269] P. Kenny, A. Hartholt, J. Gratch, D. Traum, S. Marsella, and W. Swartout, "The more the merrier: Multi-party negotiation with virtual humans," in *AAAI*, 2007.

[270] J. Laird, *The Soar Cognitive Architecture*. MIT Press, 2012.

[271] M. Sparkes, "What is a metaverse," 2021.

[272] A. Hartholt, J. Gratch, L. Weiss, and E. al., "At the virtual frontier: Introducing gunslinger, a multi-character, mixed-reality, story-driven experience," in *IVA*, 2009.

[273] M. E. Foster, A. Gaschler, M. Giuliani, A. Isard, M. Pateraki, and R. P. A. Petrick, "Two people walk into a bar: dynamic multi-party social interaction with a robot agent," in *ICMI '12*, 2012.

[274] I. Leite, M. McCoy, D. Ullman, N. Salomons, and B. Scassellati, "Comparing models of disengagement in individual and group interactions," *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 99–105, 2015.

[275] H. Candello, C. Pinhanez, M. Pichiliani, M. Guerra, and M. D. Bayser, "Having an animated coffee with a group of chatbots from the 19th century," *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[276] M. Snaith, D. G. Franco, T. Beinema, H. O. D. Akker, and A. Pease, "A dialogue game for multi-party goal-setting in health coaching," in *COMMA*, 2018.

[277] "Agentsunited." Accessed on June 30, 2023.

[278] T. Beinema, D. Davison, D. Reidsma, O. Banos, M. Bruijnes, B. Donval, Á. F. Valero, D. Heylen, D. Hofs, G. Huizing, *et al.*, "Agents united: An open platform for multi-agent conversational systems," in *21st ACM International Conference on Intelligent Virtual Agents*, 2021.

[279] K. S. J. Das, T. Beinema, H. O. D. Akker, and H. Hermens, "Generation of multi-party dialogues among embodied conversational agents to promote active living and healthy diet for subjects suffering from type 2 diabetes," in *ICT4AWE*, 2019.

[280] N. Karatas, S. Yoshikawa, and M. Okada, "Namida: Sociable driving agents with multiparty conversation," in *Proceedings of the Fourth International Conference on Human Agent Interaction*, pp. 35–42, 2016.

[281] N. Karatas, S. Yoshikawa, P. R. S. De Silva, and M. Okada, "How multi-party conversation can become an effective interface while driving," *The Transactions of Human Interface Society*, vol. 20, no. 3, pp. 371–388, 2018.

[282] S. Moubayed, G. Skantze, and J. Beskow, "The furhat back-projected humanoid head-lip reading, gaze and multi-party interaction," *Int. J. Humanoid Robotics*, vol. 10, 2013.

[283] G. Skantze and S. Moubayed, "Iristk: a statechart-based toolkit for multi-party face-to-face interaction," in *ICMI '12*, 2012.

[284] M. Johansson and G. Skantze, "Opportunities and obligations to take turns in collaborative multi-party human-robot interaction," in *SIGDIAL Conference*, 2015.

[285] G. Skantze, M. Johansson, and J. Beskow, "Exploring turn-taking cues in multi-party human-robot discussions about objects," *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015.

[286] T. Ghosal, M. Singh, A. Nedoluzhko, and O. Bojar, "Report on the sigdial 2021 special session on summarization of dialogues and multi-party meetings (summdial)," in *ACM SIGIR Forum*, no. Volume 55, Number 2, pp. 1–17, ACM New York, NY, USA, 2022.

[287] F. Yu, S. Zhang, P. Guo, Y. Fu, Z. Du, S. Zheng, W. Huang, L. Xie, Z.-H. Tan, D. Wang, *et al.*, "Summary on the icassp 2022 multi-channel multi-party meeting transcription grand challenge," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 9156–9160, IEEE, 2022.

[288] M. Purver, P. Ehlen, and J. Niekrasz, "Detecting action items in multi-party meetings: Annotation and initial experiments," in *Machine Learning for Multimodal Interaction: Third International Workshop, MLMI 2006, Bethesda, MD, USA, May 1-4, 2006, Revised Selected Papers 3*, pp. 200–211, Springer, 2006.

[289] D. M. Howcroft, A. Belz, M. A. Clinciu, D. Gkatzia, S. A. Hasan, S. Mahamood, S. Mille, E. van Miltenburg, S. Santhanam, and V. Rieser, "Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised definitions," in *INLG*, 2020.

[290] S. Gehrmann, T. P. Adewumi, K. Aggarwal, P. S. Ammanamanchi, A. Anuoluwapo, A. Bosselut, K. R. Chandu, M. A. Clinciu, D. Das, K. D. Dhole, W. Du, E. Durmus, O. Duvsek, C. C. Emezue, V. Gangal, C. Garbacea, T. Hashimoto, Y. Hou, Y. Jernite, H. Jhamtani, Y. Ji, S. Jolly, D. Kumar, F. Ladhak, A. Madaan, M. Maddela, K. Mahajan, S. Mahamood, B. P. Majumder, P. H. Martins, A. McMillan-Major, S. Mille, E. van Miltenburg, M. Nadeem, S. Narayan, V. Nikolaev, R. A. Niyongabo, S. Osei, A. P. Parikh, L. Perez-Beltrachini, N. Rao, V. Raunak, J. D. Rodríguez, S. Santhanam, J. Sedoc, T. Sellam, S. Shaikh, A. Shimorina, M. A. S. Cabezudo, H. Strobelt, N. Subramani, W. Xu, D. Yang, A. Yerukola, and J. Zhou, "The gem benchmark: Natural language generation, its evaluation and metrics," *ArXiv*, vol. abs/2102.01672, 2021.

[291] S. Kim, M. Galley, C. Gunasekara, S. Lee, A. Atkinson, B. Peng, H. Schulz, J. Gao, J. Li, M. Adada, *et al.*, "The eighth dialog system technology challenge," *arXiv preprint arXiv:1911.06394*, 2019.

[292] D. Traum, D. DeVault, J. Lee, Z. Wang, and S. Marsella, "Incremental dialogue understanding and feedback for multiparty, multimodal conversation," in *IVA*, 2012.

[293] Z. Wang, J. Lee, and S. Marsella, "Towards more comprehensive listening behavior: beyond the bobble head," in *International Workshop on Intelligent Virtual Agents*, pp. 216–227, Springer, 2011.

[294] D. Huggins-Daines, M. Kumar, A. Chan, A. W. Black, M. Ravishankar, and A. I. Rudnicky, "Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, pp. I–I, IEEE, 2006.

[295] K. Sagae, G. Christian, D. DeVault, and D. Traum, "Towards natural language understanding of partial speech recognition results in dialogue systems," in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 53–56, 2009.

[296] K. Sagae, D. DeVault, and D. Traum, "Interpretation of partial utterances in virtual human dialogue systems," in *Proceedings of the NAACL HLT 2010 Demonstration Session*, pp. 33–36, 2010.

[297] D. DeVault, K. Sagae, and D. Traum, "Incremental interpretation and prediction of utterance meaning for interactive dialogue," *Dialogue Discourse*, vol. 2, pp. 143–170, 2011.

[298] D. DeVault, K. Sagae, and D. Traum, "Detecting the status of a predictive incremental speech understanding model for real-time decision-making in a spoken dialogue system," in *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[299] D. Traum, "Semantics and pragmatics of questions and answers for dialogue agents," in *proceedings of the International Workshop on Computational Semantics*, pp. 380–394, 2003.

[300] J. Gratch and S. Marsella, "A domain-independent framework for modeling emotion," *Cognitive Systems Research*, vol. 5, no. 4, pp. 269–306, 2004.

[301] S. C. Marsella and J. Gratch, "Ema: A process model of appraisal dynamics," *Cognitive Systems Research*, vol. 10, no. 1, pp. 70–90, 2009.

[302] D. Traum and J. Rickel, "Embodied agents for multi-party dialogue in immersive virtual worlds," in *AAMAS '02*, 2002.

[303] D. Traum and L.-P. Morency, "Integration of visual perception in dialogue understanding for virtual humans in multi-party interaction," in *AAMAS 2010*, 2010.

[304] "Dialog system technology challenges 8 (dstc 8) - track 2." Accessed on June 30, 2023.

[305] J. K. Kummerfeld, S. R. Gouravajhala, J. Peper, V. Athreya, C. Gunasekara, J. Ganhotra, S. S. Patel, L. Polymenakos, and W. S. Lasecki, "A large-scale corpus for conversation disentanglement," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, July 2019.

[306] I. V. Serban, R. Lowe, P. Henderson, L. Charlin, and J. Pineau, "A survey of available corpora for building data-driven dialogue systems: The journal version," *Dialogue & Discourse*, vol. 9, no. 1, pp. 1–49, 2018.

[307] D. DiFranzo and N. N. Bazarova, "The truman platform: Social media simulation for experimental research," in *The 12th International Conference on Web and Social Media*, 2018.

[308] W. Ahmed, "Using twitter as a data source: an overview of social media research tools (updated for 2017)," *Impact of Social Sciences Blog*, 2017.

[309] "Social lab." Accessed on June 29, 2023.

[310] "Yournet." Accessed on 2018.

[311] "Fireblogger." Accessed on June 30, 2023.

[312] "Teal." Accessed on June 30, 2023.

[313] "Antisocial." Accessed on June 30, 2023.

[314] "Facebird." Accessed on June 30, 2023.

[315] "isocial." Accessed on June 30, 2023.

[316] "Friend.ly." Accessed on June 30, 2023.

[317] "bloom." Accessed on June 30, 2023.

[318] "spruce." Accessed on June 30, 2023.

[319] M. S. Granovetter, "The strength of weak ties," in *Social networks*, pp. 347–367, Elsevier, 1977.

[320] K. Chodorow and M. Dirolf, *MongoDB: The Definitive Guide*. O'Reilly Media, Inc., 1st ed., 2010.

[321] A. Mardan, *Express. js Guide: The Comprehensive Book on Express. js*. Azat Mardan, 2014.

[322] "Nodejs." Accessed on June 30, 2023.

[323] B. Green and S. Seshadri, *AngularJS*. " O'Reilly Media, Inc.", 2013.

[324] "Mean stack." Accessed on June 30, 2023.

[325] J. Cheng, L. A. Adamic, J. M. Kleinberg, and J. Leskovec, "Do cascades recur?," in *Proceedings of the 25th international conference on world wide web*, pp. 671–681, 2016.

[326] K. Lerman and R. Ghosh, "Information contagion: An empirical study of the spread of news on digg and twitter social networks," in *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[327] A. Kumar and G. Garg, "Sentiment analysis of multimodal twitter data," *Multimedia Tools and Applications*, vol. 78, no. 17, pp. 24103–24119, 2019.

[328] F. Barbieri, J. Camacho-Collados, F. Ronzano, L. E. Anke, M. Ballesteros, V. Basile, V. Patti, and H. Saggion, "Semeval 2018 task 2: Multilingual emoji prediction," in *Proceedings of The 12th International Workshop on Semantic Evaluation*, pp. 24–33, 2018.

[329] H. Almerekhi, S. b. B. J. Jansen, and c.-s. b. H. Kwak, "Investigating toxicity across multiple reddit communities, users, and moderators," in *Companion Proceedings of the Web Conference 2020*, pp. 294–298, 2020.

[330] V. A. Dounoucos, D. S. Hillygus, and C. Carlson, "The message and the medium: An experimental evaluation of the effects of twitter commentary on campaign messages," *Journal of Information Technology & Politics*, vol. 16, no. 1, pp. 66–76, 2019.

[331] J. J. Gross, "The emerging field of emotion regulation: An integrative review," *Review of general psychology*, vol. 2, no. 3, pp. 271–299, 1998.

[332] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, "Scaling instruction-finetuned language models," 2022.

[333] Y. K. Chia, P. Hong, L. Bing, and S. Poria, "Instructeval: Towards holistic evaluation of instruction-tuned large language models," *arXiv preprint arXiv:2306.04757*, 2023.

[334] R. Kumar, D. S. Chauhan, G. Dias, and A. Ekbal, "Modelling personalized dialogue generation in multi-party settings," in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6, IEEE, 2021.

[335] J. Gao, M. Galley, and L. Li, *Neural approaches to conversational AI: Question answering, task-oriented dialogues and social chatbots*. Now Foundations and Trends, 2019.

[336] A. Venkatesh, C. Khatri, A. Ram, F. Guo, R. Gabriel, A. Nagar, R. Prasad, M. Cheng, B. Hedayatnia, A. Metallinou, *et al.*, "On evaluating and comparing conversational agents," 2017.

[337] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "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)*, pp. 2204–2213, 2018.

[338] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5370–5381, 2019.

[339] H. Zhang, Z. Chan, Y. Song, D. Zhao, and R. Yan, "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*, pp. 76–84, Springer, 2018.

[340] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning, December 2014*, 2014.

[341] I. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, 2017.

[342] I. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016.

[343] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[344] D. Ju, S. Feng, P. Lv, D. Wang, and Y. Zhang, "Learning to improve persona consistency in multi-party dialogue generation via text knowledge enhancement," in *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 298–309, 2022.

[345] "Hla-chatplusplus." Accessed on July 2, 2023.

[346] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "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)*, pp. 154–164, 2019.

[347] R. Lian, M. Xie, F. Wang, J. Peng, and H. Wu, "Learning to select knowledge for response generation in dialog systems," in *International Joint Conference on Artificial Intelligence*, 2019.

[348] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *WWW'20: Proceedings of The Web Conference 2020*, 2020.

[349] "Hetermpc." Accessed on July 2, 2023.

[350] M. Wang, D. Zheng, Z. Ye, Q. Gan, M. Li, X. Song, J. Zhou, C. Ma, L. Yu, Y. Gai, *et al.*, "Deep graph library: A graph-centric, highly-performant package for graph neural networks," *arXiv preprint arXiv:1909.01315*, 2019.

[351] Y. Sun, J. Han, X. Yan, P. S. Yu, and T. Wu, "Pathsim: Meta path-based top-k similarity search in heterogeneous information networks," *Proceedings of the VLDB Endowment*, vol. 4, no. 11, pp. 992–1003, 2011.

[352] Y. Sun, B. Norick, J. Han, X. Yan, P. S. Yu, and X. Yu, "Pathselclus: Integrating meta-path selection with user-guided object clustering in heterogeneous information networks," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 7, no. 3, pp. 1–23, 2013.

[353] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *International Conference on Learning Representations*, 2016.

[354] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[355] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim, "Graph transformer networks," *Advances in neural information processing systems*, vol. 32, 2019.

[356] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "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*, pp. 593–607, Springer, 2018.

[357] X. Wang, H. Ji, C. Shi, B. Wang, Y. Ye, P. Cui, and P. S. Yu, "Heterogeneous graph attention network," in *The world wide web conference*, pp. 2022–2032, 2019.

[358] C. Zhang, D. Song, C. Huang, A. Swami, and N. V. Chawla, "Heterogeneous graph neural network," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 793–803, 2019.

[359] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[360] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

[361] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017.

[362] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[363] D. M. Howcroft, A. Belz, M.-A. Clinciu, D. Gkatzia, S. A. Hasan, S. Mahamood, S. Mille, E. van Miltenburg, S. Santhanam, and V. Rieser, "Twenty years of confusion in human evaluation: Nlg needs evaluation sheets and standardised

definitions," in *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 169–182, 2020.

[364] A. Belz, S. Mille, and D. M. Howcroft, "Disentangling the properties of human evaluation methods: A classification system to support comparability, meta-evaluation and reproducibility testing," in *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 183–194, 2020.

[365] E. Clark, T. August, S. Serrano, N. Haduong, S. Gururangan, and N. A. Smith, "All that's 'human' is not gold: Evaluating human evaluation of generated text," 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)*, pp. 7282–7296, 2021.

[366] M. Hämäläinen and K. Alnajjar, "The great misalignment problem in human evaluation of nlp methods," in *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pp. 69–74, 2021.

[367] C. van der Lee, A. Gatt, E. van Miltenburg, and E. Krahmer, "Human evaluation of automatically generated text: Current trends and best practice guidelines," *Computer Speech & Language*, vol. 67, p. 101151, 2021.

[368] S. Gehrmann, T. Adewumi, K. Aggarwal, P. S. Ammanamanchi, A. Anuoluwapo, A. Bosselut, K. R. Chandu, M. Clinciu, D. Das, K. D. Dhole, *et al.*, "The gem benchmark: Natural language generation, its evaluation and metrics," *arXiv preprint arXiv:2102.01672*, 2021.

[369] D. Khashabi, G. Stanovsky, J. Bragg, N. Lourie, J. Kasai, Y. Choi, N. A. Smith, and D. S. Weld, "Genie: A leaderboard for human-in-the-loop evaluation of text generation," *arXiv preprint arXiv:2101.06561*, 2021.

[370] P. Liu, J. Fu, Y. Xiao, W. Yuan, S. Chang, J. Dai, Y. Liu, Z. Ye, and G. Neubig, "Explainaboard: An explainable leaderboard for nlp," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 280–289, 2021.

[371] S. Mille, K. Dhole, S. Mahamood, L. Perez-Beltrachini, V. Gangal, M. Kale, E. van Miltenburg, and S. Gehrmann, "Automatic construction of evaluation suites for natural language generation datasets," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

[372] R. Higashinaka, K. Funakoshi, M. Araki, H. Tsukahara, Y. Kobayashi, and M. Mizukami, "Towards taxonomy of errors in chat-oriented dialogue systems," in *Proceedings of the 16th annual meeting of the special interest group on discourse and dialogue*, pp. 87–95, 2015.

[373] R. Higashinaka, M. Araki, H. Tsukahara, and M. Mizukami, "Improving taxonomy of errors in chat-oriented dialogue systems," in *9th International Workshop on Spoken Dialogue System Technology*, pp. 331–343, Springer, 2019.

[374] A. Belz and E. Reiter, "Comparing automatic and human evaluation of nlg systems," in *11th conference of the european chapter of the association for computational linguistics*, pp. 313–320, 2006.

[375] E. Reiter and A. Belz, "An investigation into the validity of some metrics for automatically evaluating natural language generation systems," *Computational Linguistics*, vol. 35, no. 4, pp. 529–558, 2009.

[376] J. Novikova, O. Dušek, A. C. Curry, and V. Rieser, "Why we need new evaluation metrics for nlg," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2241–2252, 2017.

[377] S. Santhanam and S. Shaikh, "Towards best experiment design for evaluating dialogue system output," in *Proceedings of the 12th International Conference on Natural Language Generation*, (Tokyo, Japan), pp. 88–94, Association for Computational Linguistics, Oct.–Nov. 2019.

[378] S. Santhanam, A. Karduni, and S. Shaikh, "Studying the effects of cognitive biases in evaluation of conversational agents," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2020.

[379] D. Scott and J. Moore, "An nlg evaluation competition? eight reasons to be cautious," in *Proceedings of the Workshop on Shared Tasks and Comparative Evaluation in Natural Language Generation*, pp. 22–23, 2007.

[380] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[381] "Parlai." Accessed on June 30, 2023.

[382] S. Banerjee and A. Lavie, "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*, pp. 65–72, 2005.

[383] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, pp. 74–81, 2004.

[384] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2122–2132, 2016.