Who Says What to Whom: A Survey of Multi-Party Conversations

Jia-Chen Gu[1], **Chongyang Tao**[2] and **Zhen-Hua Ling**[1]

[1]National Engineering Research Center for Speech and Language Information Processing,
University of Science and Technology of China, Hefei, China
[2]Microsoft, Beijing, China
gujc@mail.ustc.edu.cn, chotao@microsoft.com, zhling@ustc.edu.cn

Abstract

Multi-party conversations (MPCs) are a more practical and challenging scenario involving more than two interlocutors. This research topic has drawn significant attention from both academia and industry, and it is nowadays counted as one of the most promising research areas in the field of dialogue systems. In general, MPC algorithms aim at addressing the issues of *Who* says *What* to *Whom*, specifically, who speaks, say what, and address whom. The complicated interactions between interlocutors, between utterances, and between interlocutors and utterances develop many variant tasks of MPCs worth investigation. In this paper, we present a comprehensive survey of recent advances in text-based MPCs. In particular, we first summarize recent advances on the research of MPC context modeling including dialogue discourse parsing, dialogue flow modeling and self-supervised training for MPCs. Then we review the state-of-the-art models categorized by *Who* says *What* to *Whom* in MPCs. Finally, we highlight the challenges which are not yet well addressed in MPCs and present future research directions.

1 Introduction

Enabling dialogue systems to converse naturally with humans is a challenging yet intriguing problem of artificial intelligence, which has attracted increasing attention due to its promising potentials and alluring commercial values in building dialogue agents [Zhou *et al.*, 2020]. Most of the existing methods focus on building dialogue systems between two interlocutors, commonly known as two-party conversations (TPCs) [Serban *et al.*, 2016; Wu *et al.*, 2017; Zhou *et al.*, 2018; Zhang *et al.*, 2020]. Recently, researchers have paid more attention to a more practical and challenging scenario involving more than two interlocutors, commonly known as multi-party conversations (MPCs) [Traum, 2004; Uthus and Aha, 2013; Ouchi and Tsuboi, 2016; Mehri and Carenini, 2017; Zhang *et al.*, 2018; Hu *et al.*, 2019; Wang *et al.*, 2020; Gu *et al.*, 2021].

Utterances in a TPC are posted between two interlocutors alternately, constituting a *sequential* information flow. On the other hand, each utterance in an MPC can be spoken by anyone and address anyone else in this conversation, which constitutes a *graphical* information flow as shown in Figure 1. The complicated interactions between interlocutors, between utterances, and between interlocutors and utterances develop many variant tasks of MPCs worth investigation. For example, there are scenarios where multiple people may want to interact with a chatbot or one person interacts with several chatbots at the same time, as in a chat group, to coordinate among themselves and achieve a common goal. Thus, the ability to predict which agent in the conversation is the most likely to speak next, and conversely, when an agent must wait before interacting, is important for conducting engaging and social conversations [Pinhanez *et al.*, 2018; de Bayser *et al.*, 2019]. Furthermore, detecting who is being addressed, i.e., who the current speaker is talking to, is also non-trivial in these conversation scenarios [Ouchi and Tsuboi, 2016; Zhang *et al.*, 2018; Le *et al.*, 2019; Gu *et al.*, 2021]. Last but not least, only after knowing a speaker and an addressee at the current dialogue state, can the system return an appropriate response following the conversation structure [Zhang *et al.*, 2018; Hu *et al.*, 2019; Wang *et al.*, 2020; Gu *et al.*, 2021].

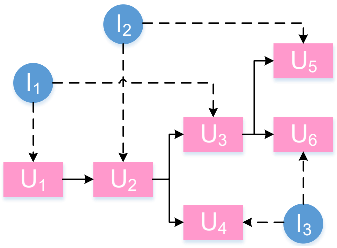

Figure 1: Illustration of a graphical information flow in an MPC. Pink rectangles denote utterances and blue circles denote interlocutors. Each solid line represents the relationship that an utterance is replied by another one. Each dashed line indicates the speaker of an utterance.

In contrast to the prosperity of research surveys in TPCs [Chen *et al.*, 2017; Huang *et al.*, 2020; Tao *et al.*, 2021], to the best of our knowledge, there are no updated and systematic introductions to tasks and methods in MPCs after Traum, 2004, to summarize recent advances in MPCs. Thus, in this survey paper, we present a literature

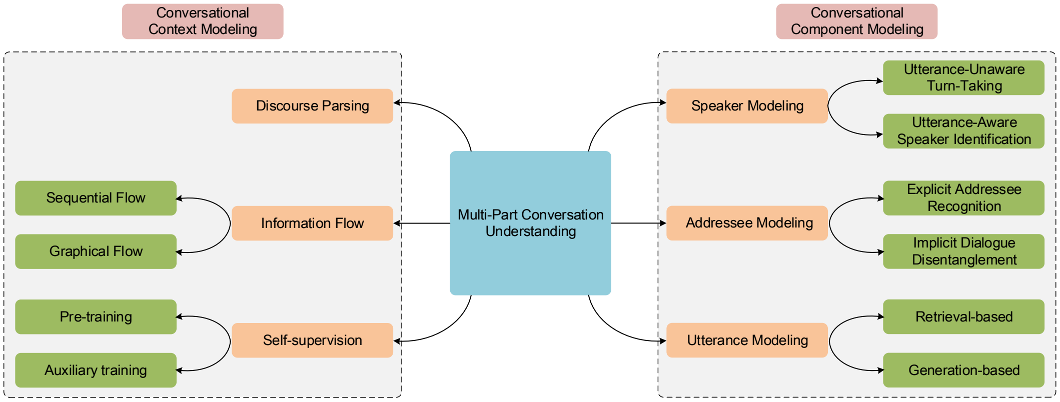
Conversational Context Modeling
Conversational Component Modeling
Discourse Parsing
Sequential Flow
Information Flow
Graphical Flow
Pre-training
Self-supervision
Auxiliary training
Multi-Part Conversation Understanding
Speaker Modeling
Utterance-Unaware Turn-Taking
Utterance-Aware Speaker Identification
Addressee Modeling
Explicit Addressee Recognition
Implicit Dialogue Disentanglement
Utterance Modeling
Retrieval-based
Generation-based

Figure 2: The taxonomy of studies in MPCs, categorized into *conversational context modeling* and *conversational component modeling*.

review of recent advances for tasks and their corresponding methods in text-based MPCs, and organize these studies into a taxonomy. Note that review on MPC datasets is not the focus of this survey, thus we recommend readers to refer to Mahajan and Shaikh for full informativeness. As shown in Figure 2, we organize the existing studies into two categories of *conversational context modeling* and *conversational component modeling*, according to the core of task evaluation. The rest of this survey is organized as follows. It starts with a summary of recent advances on text-based MPC context modeling including dialogue discourse parsing, dialogue flow modeling and self-supervised training techniques for MPCs. Then, we review the state-of-the-art (SOTA) models on MPC component modeling that study *Who* says *What* to *Whom* basically. Finally, based on analysis of the existing studies, we highlight open challenges that are not yet well addressed and present future research directions.

In summary, our contributions in this paper are three-fold: (1) To the best of our knowledge, this paper presents an updated and comprehensive survey on recent advances in multi-party conversations. (2) We review the existing studies thoroughly and organize them into an appropriate taxonomy. (3) Several open challenges are summarized and highlighted to facilitate the research community for future work.

2 Conversational Context Modeling

It is important to model a conversation effectively and efficiently for better MPC understanding among both high-level and basic tasks in MPCs. In this section, we summarize the studies on conversational context modeling and categorize them into dialogue discourse parsing, dialogue information flow and self-supervised training for MPCs.

2.1 Discourse Parsing

The *reply-to* relation [Asher *et al.*, 2016] is a general conception to describe all possible relations between a pair of utterances. The *reply-to* relation can be further specified to many types of relations between any pair of utterances in detail. For example, a follow-up utterance can be a *clarification* or an *answer* of a question raised in the conversation history, or an *acknowledgement* or an *elaboration* of a comment [Asher *et al.*, 2016]. Thus, the task of discourse parsing in MPCs is designed to analyze the discourse structure to specify inter-utterance relations at a fine granularity [Afantenos *et al.*, 2015; Asher *et al.*, 2016; Perret *et al.*, 2016; Shi and Huang, 2019; Wang *et al.*, 2021].

Afantenos *et al.* and Asher *et al.* first define the task formalization of discourse parsing in MPCs, and describe the STAC corpus collected from an online version of the game *The Settlers of Catan*. Perret *et al.* incorporate integer linear programming (ILP) to encode both objective functions and constraints as global decoding over local scores. In this way, the interaction between a pair of utterances and the global information over the whole dialogue can be both used for predicting a relation. Shi and Huang propose a neural model to predict dependency relations and construct discourse structures jointly and alternately. They further utilize the local information that encodes the concerned utterances, the global information that encodes the currently constructed structure, and the speaker information for enhancing conversation understanding. Though using previously predicted structure can provide richer information for constructing structures, it can also lead to problems with severe error propagation. To alleviate error propagation, Wang *et al.* adopt an edge-centric graph neural network to update the information between each utterance pair layer by layer, so that expressive representations can be learned without historical predictions. This approach achieves SOTA performance on MPC discourse parsing and readers can start from here [1].

2.2 Information Flow

Information flows in MPCs are generally modeled as either *sequential* or *graphical*.

Sequential Information Flow Studies on TPCs model a multi-turn session as a concatenation [Lowe *et al.*, 2015; Zhang *et al.*, 2020] or a sequence of utterances [Serban *et al.*, 2016; Wu *et al.*, 2017; Zhou *et al.*, 2018]. At the beginning, most of studies on building MPC systems follow this sequential flow framework [Ouchi and Tsuboi, 2016; Zhang *et al.*, 2018; Shi and Huang, 2019; Wang *et al.*, 2020]. Although these methods are effective in modeling the sequential information in MPCs to some extent, they fall

[1] https://github.com/Soistesimmer/Structure-Self-Aware

short in modeling complicated *reply-to* relationships within. For example, in Figure 1, U_3 and U_4 both reply to U_2, while U_5 and U_6 both reply to U_3. How to encode this type of parallel relationship is beyond the representation capability of sequential-flow-based models.

Graphical Information Flow To overcome the deficiencies in the ability to represent this type of structural information, recent studies have instead modeled MPC information flows with a graph topology. Hu *et al.* propose to encode utterances based on a homogeneous graph rather than the sequence of their appearances for response generation. Gu *et al.* improve this method to take both utterances and interlocutors into consideration with a heterogeneous graph, and to model heterogeneity by introducing node-edge-type-dependent parameters. Ghosal *et al.* propose to use a directed graph for emotion recognition, the nodes of which represent utterances, and the edges represent the dependency between those utterance speakers along with their relative positions in a conversation. Then this graph is fed to a graph convolution network (GCN). Jia *et al.* make use of an additional labeled dataset for training a dependency parser, which is utilized to uncover the graphical dependency relations and to disentangle the parallel threads into self-contained sub-conversations. Feng *et al.* and Li *et al.* propose to incorporate discourse relations between utterances, and convert context utterances to a discourse-aware dialogue graph for meeting summarization and reading comprehension respectively. Lee and Choi construct a heterogeneous dialogue graph to capture the relational information in a dialogue, which consists of four types of nodes for dialogue, turn, subject and object, along with three types of edge for speaker, dialogue and argument. This approach achieves SOTA performance on MPC graph representation and readers can start from here [2].

2.3 Self-supervision

Since an MPC instance always contains complicated interactions between interlocutors, between utterances, and between interlocutors and utterances, it is beneficial to make use of these complementary self-supervised signals for better conversation understanding. Recently, researchers have been paying attention to designing various self-supervised tasks, so that models can produce better contextualized interlocutor and utterance representations which can be effectively generalized to multiple downstream tasks of MPCs.

Pre-training A line of research follows the two-stage framework of pre-training and fine-tuning. Wu *et al.*; Bertero *et al.* and Gu *et al.* continue to post-train the pre-trained language models (PLMs) with the self-supervised tasks of mask language model (MLM) and next sentence prediction (NSP) to incorporate domain knowledge. After that, models are fine-tuned for the downstream task of response selection. To learn better representations that can be generalized to more downstream tasks, Gu *et al.* design two major types of self-supervised tasks for modeling interlocutor structures and utterance semantics in a unified framework for universal MPC understanding. Then, models are fine-tuned on different downstream tasks of addressee recognition, speaker identification and response selection for evaluating the generalization ability of the pre-trained model. This approach achieves SOTA performance on MPC pre-training and readers can start from here [3].

Auxiliary Training The other line of research designs auxiliary tasks that are employed to train models along with a main task. Wang *et al.*, 2020 design two auxiliary tasks of topic prediction and topic disentanglement to equip PLMs with the ability of tracking parallel topics in a MPC for response selection. Wang *et al.*, 2021 introduce two auxiliary tasks of relation recognition and structure distillation for better representation learning in discourse parsing. Li and Zhao design two auxiliary tasks for conversational reading comprehension. One is a self-supervised speaker prediction task to implicitly model speaker information flows, and the other is a pseudo-self-supervised key-utterance prediction task to capture salient utterances in a long and noisy dialogue. These auxiliary tasks are designed for respective tasks and readers can take inspiration from these methods for reference.

3 Conversational Component Modeling

On top of a summary of the complicated interactions between interlocutors, between utterances, and between interlocutors and utterances, the underlying challenge lies in solving the issue of *Who* says *What* to *Whom* basically. The three components of who speaking, saying what, and addressing whom are further introduced, since a (*speaker*, *utterance*, *addressee*) triple constitutes a basic unit of an MPC.

3.1 WHO Speaks

An important issue in MPCs is speaker modeling, which aims at determining the next speaker or the speaker of a specific utterance in a conversation.

Utterance-Unaware Turn-Taking *Turn-taking*, is a task that attempts to determine which speaker has the floor after each utterance, given the speakers and what they said in the previous conversation history [Hawes *et al.*, 2009; Pinhanez *et al.*, 2018; de Bayser *et al.*, 2018; de Bayser *et al.*, 2019]. Formally, given speakers and utterances of the previous $n-1$ conversation turns, i.e., $\{(s_1, u_1), ..., (s_{n-1}, u_{n-1})\}$ where s_i and u_i denote a speaker and an utterance of the i-th turn respectively, models are asked to predict the speaker of the n-th turn, i.e., s_n. This task involves modeling the coordination strategies that speakers adopt to acquire or give up the floor, so that an ongoing conversation can go on smoothly. Hawes *et al.* treat this task as a supervised sequence-labeling problem and use the first- and second-order conditional random fields (CRFs) model to predict the next speaker in Supreme Court oral argument transcripts. Pinhanez *et al.* and de Bayser *et al.* [2018] test an MPC system named *finch* that enables interactions between one person and four collaborative chatbots which are experts in financial investments. In this system, turn-taking is controlled by a rule-based Finite-State Automata (FSA) service. This service is called for every utterance exchanged in the group chat by considering both conversation history and interactions between interlocutors.

[2] https://github.com/BlackNoodle/TUCORE-GCN

[3] https://github.com/JasonForJoy/MPC-BERT

de Bayser *et al.* [2019] present comparisons of employing different machine learning techniques such as maximum likelihood estimation (MLE), support vector machines (SVM), convolutional neural networks (CNN) and long short-term memory (LSTM) to predict the next speaker.

Utterance-Aware Speaker Identification On the other hand, who will be the next speaker in an MPC could have multiple answers since multiple sub-conversation threads are assumed to occur simultaneously. The result of speaker identification may vary when considering the posterior information of different utterance semantics. Thus, researchers are paying more attention to utterance-aware speaker modeling conditioned on the semantics of an utterance to predict [Glass and Bangay, 2007; Elson and McKeown, 2010; O'Keefe *et al.*, 2012; Iosif and Mishra, 2014; Meng *et al.*, 2017; Meng *et al.*, 2018; Chen *et al.*, 2019; Gu *et al.*, 2021].

A line of research studies the task of *speaker identification in English novels* for determining who says a quote in a given context by text analysis. This task is based exclusively on conversations extracted from the novels. These studies can be generally categorized into rule-based [Glass and Bangay, 2007; O'Keefe *et al.*, 2012; He *et al.*, 2013] and machine-learning-based methods [Elson and McKeown, 2010; Pareti *et al.*, 2013; Iosif and Mishra, 2014]. Afterwards, Chen *et al.*, 2019 present a Chinese dataset for identifying speakers in novels, along with a rule-based and a classifier-based baselines. Chen *et al.* 2021 improve by formulating it as a scoring task and designing a BERT-based scoring network.

There are also other utterance-aware speaker-related modeling tasks. The task of *speaker segmentation*, also known as *speaker diarisation* or *speaker change detection*, aims at finding speaker changing points in a conversation. Specifically, a speaker change occurs when the current and the previous utterances are not uttered by the same speaker. Meng *et al.*, 2017 first propose text-based speaker segmentation by formulating this task as a binary utterance-pair classification task to judge whether the speaker is changing before and after a certain decision point. The semantics discrepancy of the utterance-pair play an important role to this task. Meng *et al.*, 2018 propose another task of speaker classification as a surrogate for general speaker modeling. This task is defined by segmenting an MPC into several parts according to speakers, each segment of which comprises one or a few consecutive sentences uttered by a particular speaker. A candidate set of speakers is also given and models are required to identify the speaker of each segment. Gu *et al.* propose an utterance semantics-based speaker searching task where models are asked to search for a speaker in conversation history that shares the same speaker with the n-th conversation turn, given not only speakers and utterances of the previous $n-1$ turns, but also the utterance of the n-th turn. This approach achieves SOTA performance on MPC speaker identification and readers can start from here [3].

3.2 Say WHAT

Different from modeling speakers and addressees that are unique issues in MPCs, the issue of *saying what* is common in both TPCs and MPCs. Similar to studies on TPCs, existing methods enabling dialogue systems to decide what to say in MPCs can be generally categorized into *retrieval*-based [Ouchi and Tsuboi, 2016; Zhang *et al.*, 2018; Wu *et al.*, 2020; Wang *et al.*, 2020; Gu *et al.*, 2021] and *generation*-based methods [Liu *et al.*, 2019; Hu *et al.*, 2019]. In addition to modeling semantics, consistency, and interactiveness [Huang *et al.*, 2020] between contexts and responses in TPCs [Lowe *et al.*, 2015; Serban *et al.*, 2016; Wu *et al.*, 2017; Zhou *et al.*, 2018; Tao *et al.*, 2019; Zhang *et al.*, 2020], the complex conversation structures and topic transitions also contribute to the performance in MPCs. We discuss what should be said as a conversational agent in retrieval- and generation-based manners.

Retrieval-based The retrieval-based methods aims at selecting the best-matched response from a set of candidates, given the context of a multi-turn conversation. The key to this task is to rank the set of response candidates according to the semantic matching between the context of an MPC and a response candidate [Ouchi and Tsuboi, 2016; Zhang *et al.*, 2018; Wu *et al.*, 2020; Wang *et al.*, 2020; Gu *et al.*, 2021]. Ouchi and Tsuboi propose jointly modeling the tasks of response selection and addressee selection to capture what is being said to whom at each time step in a context. Zhang *et al.* follow this framework and improve it by updating the interlocutor embeddings role-sensitively. Gu *et al.* propose jointly learning who says what to whom in a unified framework by considering addressee- and speaker-related tasks as complementary signals for response selection during the pre-training stage.

Another line of research for response selection in MPCs is proposed in the Eighth Dialog System Technology Challenge (DSTC8) [Kim *et al.*, 2021] that extends the previous one in DSTC7 by considering full excerpt conversations in the IRC channel. Since a full excerpt is composed of hundreds of utterances, it is necessary to pre-filter these utterances for condensing and refining the conversation history. Wu *et al.*; Bertero *et al.* and Gu *et al.* rank the top three positions in this track respectively that all design various heuristic rules to extract addressee- and speaker-related features for context disentanglement and selecting relevant utterances, along with the techniques of data augmentation, post-training, and ensembling, to improve performance of response selection. Wang *et al.* outperform previous work by framing this task as dynamic topic tracking with the intuition that topic should remain the same as going from contexts to responses. It achieves SOTA performance on MPC response selection and readers can start from here [4].

Generation-based The generation-based methods synthesize a response with a natural language generation model by maximizing its generation probability given the previous conversation history [Liu *et al.*, 2019; Hu *et al.*, 2019]. Liu *et al.* incorporate interlocutor-aware contexts into recurrent encoder-decoder frameworks (ICRED) by interacting to capture contexts for different interlocutors, and leveraging an addressee memory to enhance the contextual interlocutor information for a target addressee. Hu *et al.* propose a graph-structured neural network (GSN), the core of which

[4] https://github.com/salesforce/TopicBERT

is to encode utterances based on the graph topology rather than the sequence of their appearances in a conversation, in order to model the information flow as *graphical*. Gu *et al.* propose to model complicated interactions between utterances and interlocutors in MPCs with a heterogeneous graph, where two types of graph nodes and six types of edges are designed to model heterogeneity. This approach achieves SOTA performance on MPC response generation and readers can start from here [5].

3.3 Address WHOM

Detecting who is being addressed, i.e., a behavior whereby interlocutors indicate to whom they are speaking, is non-trivial in MPCs. Existing methods on addressee modeling can be generally categorized into *explicit* and *implicit* ones.

Explicit Addressee Modeling: Addressee Recognition The task of *explicitly* determining the intended recipient of an utterance in a conversation is called *addressee recognition* [Ouchi and Tsuboi, 2016; Zhang *et al.*, 2018; Le *et al.*, 2019; Gu *et al.*, 2021]. Different from face-to-face conversations, the explicit declaration of names of addressees is more common in text-only-based conversations. Previous studies mainly focus on predicting the addressee of only the last utterance of a conversation [Ouchi and Tsuboi, 2016; Zhang *et al.*, 2018], while recent studies pay more attention to predicting the addressees of all utterances of a conversation [Le *et al.*, 2019; Gu *et al.*, 2021]. Ouchi and Tsuboi propose two frameworks of static and dynamic modeling, and show that the recurrent neural network-based models of these two frameworks predict the addressee of the last utterance of a conversation robustly. Zhang *et al.* improve the dynamic modeling framework by distinguishing the interlocutor roles (sender, addressee, observer) at a finer granularity, and updating the interlocutor embeddings role-sensitively, since interlocutors play one of the three roles at each turn and those roles vary across turns. Le *et al.* propose a who-to-whom (W2W) model to recognize and complete the addressees of all utterances in a conversation to help understand the whole conversation, given an MPC where part of the addressees are unspecified. Gu *et al.* propose a pre-trained MPC-BERT language model for universal MPC understanding by designing self-supervised tasks, and test it on addressee recognition. This approach achieves SOTA performance on addressee recognition and readers can start from here [3].

Implicit Addressee Modeling: Dialogue Disentanglement When multiple conversations occur simultaneously, a listener must decide which conversation each utterance is part of in order to interpret and respond to it appropriately. This task is referred as *dialogue disentanglement* or *thread detection* [Shen *et al.*, 2006; Elsner and Charniak, 2008; Elsner and Schudy, 2009; Elsner and Charniak, 2010; Elsner and Charniak, 2011; Riou *et al.*, 2015; Mehri and Carenini, 2017; Jiang *et al.*, 2018; Kummerfeld *et al.*, 2019; Li *et al.*, 2020b; Liu *et al.*, 2020; Jiang *et al.*, 2021]. The messages from different interlocutors on different topics are heavily interwoven. Therefore, it is necessary to disentangle a whole conversation into several threads from a data stream so that each thread is about a specific topic. Basically, most of the existing methods are designed to find out which previous utterance in the history the current utterance is replying to. Thus, this task is in essence modeling addressees *implicitly*.

Most of previous studies evaluate their methods on datasets collected by their own that are usually not publicly released and cannot provide fair comparisons between methods. Elsner and Charniak present a corpus collected from the `#Linux` internet relay chat (IRC) channel in which various conversations have been manually disentangled. This `#Linux` IRC dataset is publicly available and facilitates a line of research for improvements of disentanglement. Elsner and Schudy evaluate a variety of heuristic solvers for correlation clustering. Wang and Oard exploit the inter-dependency between the meaning of a message along with its temporal and social contexts to provide a more accurate message representation. Elsner and Charniak extend their graph-theoretic clustering model with two strategies of specificity tuning and conversation start detection. Elsner and Charniak demonstrate that several popular models focusing on local discourse coherence transfer well to the task of disentanglement. Riou *et al.* adapt the system in Elsner and Charniak to French language chats by annotating conversations and discourse relations in the `#Ubuntu-fr` channel. It also finds that using discursive information, in the form of functional and rhetoric relations between messages, is valuable for this task. Mehri and Carenini re-annotate the reply-to structure of a subset of the `#Linux` IRC dataset, and explore the usage of a reply classifier and an RNN to learn semantic relationships between messages. Jiang *et al.* propose a two-stage approach consisting of message pair similarity estimation that integrates local and global representations of messages, and conversation identification that ranks messages within a time window and constructs a message graph. Tan *et al.* combine two context-aware thread detection models that capture contexts of existing threads and conversational flows.

A remarkable effort in this direction is the work of Kummerfeld *et al.*, which creates a large-scale `#Ubuntu` IRC corpus that is 16 times larger than all previously released datasets combined. It also proposes a feed-forward neural network using additional features such as time, directedness, word overlap, and context. This corpus is first used as a benchmark in a track of DSTC8 [Kim *et al.*, 2021]. Gu *et al.* rank first in this track by employing BERT [Devlin *et al.*, 2019] to capture the matching information in each utterance pair at utterance-level, and a BiLSTM to aggregate context-level semantics. Based on this corpus, Zhu *et al.* design a masking mechanism denoting which history utterances are attendable for a target utterance, to learn conversation structures, aggregate ancestors and guide the ancestor flow. Pappadopulo *et al.* apply the Directed-Acyclic-Graph LSTM to this task for a systematic inclusion of the structured information, e.g., user turns and mentions, in the learned representation of the conversation context. Yu and Joty formalize the link prediction of disentanglement as a pointing problem using pointer networks [Vinyals *et al.*, 2015]. Specifically, each pointing operation is modeled as a multinomial distribution over the set of previous utterances,

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

which achieves SOTA performance on MPC disentanglement and readers can start from here [6].

There are also other studies that apply disentanglement to specific scenarios. Liu *et al.*, 2020 and Liu *et al.*, 2021 study end-to-end transition-based model for online dialogue disentanglement. Liu *et al.*, 2020 model the semantic coherence within each session by sequentially adding utterances into their best-matching sessions. Liu *et al.*, 2021 further explore unsupervised disentanglement upon a deep co-training algorithm that optimizes the message-pair and session classification through reinforcement learning. Jiang *et al.* investigate how well the existing disentanglement measures reflect human satisfaction in the domain of software engineering by leveraging Levenshtein distance and ratio.

4 Conclusion and Open Challenges

This paper reviews recent studies for text-based MPCs and organizes them in a novel taxonomy, including conversational context modeling and conversational component modeling. Although extensive efforts have been made and impressive results have been achieved on many benchmarks of MPCs, there are still several open challenges.

Universal MPC Understanding Most of existing studies design models for each individual task in MPCs separately. Intuitively, the complicated interactions between interlocutors, between utterances, and between interlocutors and utterances might make these tasks complementary among each other. Making use of these tasks simultaneously may produce better contextualized representations of interlocutors and utterances, and would enhance the conversation understanding. Gu *et al.* first propose several supervised tasks for pre-training a language model focusing on modeling only interlocutor structures and utterance semantics. A future research direction could be designing better supervised tasks for equipping PLMs with more abilities (such as discourse parsing), and being tested on more downstream tasks of MPCs to evaluate the robustness and generalization of a model. Furthermore, an MPC system should be able to track topic transitions and update the conversation structure dynamically as a conversation proceeds, so that it can understand deep semantics and structures of an MPC. Wang *et al.* maintain a general topic kernel and to check whether the topic remains consistent between contexts and responses. Track the specific flows of multiple ongoing sub-conversations at a fine granularity is also worth investigation, since a shared topic tracker at a coarse granularity cannot understand in-depth.

Modeling Heterogeneity in MPCs There has been a trend to model MPCs with a graph. They usually employ a homogeneous graph to cover only interactions between utterances [Hu *et al.*, 2019; Ghosal *et al.*, 2019; Feng *et al.*, 2021; Li *et al.*, 2021]. However, there are also other important components such as speakers or addressees in an MPC. To model a wider range of interactions, it is necessary to put different components in a unified graph. Thus, it is a promising direction to construct a heterogeneous graph for covering different components in MPCs, and establish relation-type dependent edges for maximizing the feature distribution differences. The complicated relationships between various components might be well modeled.

High-level MPC Application Different from the conventional studies on single-party document understanding, such as summarization [Cohan *et al.*, 2018] and machine reading comprehension [Rajpurkar *et al.*, 2018], it is challenging to understand high-level MPC applications, since the key messages of an MPC are often scattered, spanning multiple utterances from different interlocutors and leading to low information density. Furthermore, topic drifts, frequent coreferences, diverse interactive signals and domain terminologies increase the difficulty of this task. Currently, most of the existing studies work on a specific domain, such as meeting, chat, email thread, customer service, and medical dialogue [Feng *et al.*, 2022]. However, the input of dialogues differs greatly according to domains, since interlocutors may play different roles in each domain with different formats of verbal expressions. Thus, it is beneficial to construct a high-level MPC understanding system showing great compatibility across domains.

Low- or Zero-resource MPC Modeling As we mentioned above that there are many scenarios involving MPC modeling, and the cost of annotating datasets may vary by domain. Thus, it is difficult to train such complex models under low- or zero-resource settings. Researchers have studied this issue in TPCs [Zhao *et al.*, 2019; Li *et al.*, 2020a], which has not yet been explored in MPCs. If this can be done, one can transfer the ability of MPC modeling from datasets that are easily accessible to those are hard to obtain, showing great performance under low- or zero-resource scenarios.

References

[Afantenos *et al.*, 2015] Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. Discourse parsing for multi-party chat dialogues. In *EMNLP*, 2015.

[Asher *et al.*, 2016] Nicholas Asher, Julie Hunter, Mathieu Morey, et al. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *LREC*, 2016.

[Bertero *et al.*, 2020] Dario Bertero, Takeshi Homma, Kenichi Yokote, et al. Model ensembling of ESIM and BERT for dialogue response selection. In *DSTC8*, 2020.

[Chen *et al.*, 2017] Hongshen Chen, Xiaorui Liu, Dawei Yin, et al. A survey on dialogue systems: Recent advances and new frontiers. *SIGKDD Explor. Newsl.*, 2017.

[Chen *et al.*, 2019] Jia-Xiang Chen, Zhen-Hua Ling, and Li-Rong Dai. A Chinese dataset for identifying speakers in novels. In *INTERSPEECH*, 2019.

[Chen *et al.*, 2021] Yue Chen, Zhen-Hua Ling, and Qing-Feng Liu. A neural-network-based approach to identifying speakers in novels. In *INTERSPEECH*, 2021.

[Cohan *et al.*, 2018] Arman Cohan, Franck Dernoncourt, et al. A discourse-aware attention model for abstractive summarization of long documents. In *NAACL*, 2018.

[de Bayser *et al.*, 2018] Maira Gatti de Bayser, Melina Alberio Guerra, Paulo Cavalin, et al. Specifying and

[6] https://github.com/vode/onlinePtrNet_disentanglement

implementing multi-party conversation rules with finite-state-automata. In *AAAI Workshop*, 2018.

[de Bayser *et al.*, 2019] Maira Gatti de Bayser, Paulo Cavalin, Claudio Pinhanez, et al. Learning multi-party turn-taking models from dialogue logs. 2019.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, et al. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019.

[Elsner and Charniak, 2008] Micha Elsner and Eugene Charniak. You talking to me? a corpus and algorithm for conversation disentanglement. In *ACL*, 2008.

[Elsner and Charniak, 2010] Micha Elsner and Eugene Charniak. Disentangling chat. *CL*, 2010.

[Elsner and Charniak, 2011] Micha Elsner and Eugene Charniak. Disentangling chat with local coherence models. In *ACL*, 2011.

[Elsner and Schudy, 2009] Micha Elsner and Warren Schudy. Bounding and comparing methods for correlation clustering beyond ilp. In *ILPNLP*, 2009.

[Elson and McKeown, 2010] David K Elson and Kathleen R McKeown. Automatic attribution of quoted speech in literary narrative. In *AAAI*, 2010.

[Feng *et al.*, 2021] Xiachong Feng, Xiaocheng Feng, Bing Qin, et al. Dialogue discourse-aware graph model and data augmentation for meeting summarization. In *IJCAI*, 2021.

[Feng *et al.*, 2022] Xiachong Feng, Xiaocheng Feng, and Bing Qin. A survey on dialogue summarization: Recent advances and new frontiers. In *IJCAI*, 2022.

[Ghosal *et al.*, 2019] Deepanway Ghosal, Navonil Majumder, et al. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *EMNLP*, 2019.

[Glass and Bangay, 2007] Kevin Glass and Shaun Bangay. A naive salience-based method for speaker identification in fiction books. In *PRASA*, 2007.

[Gu *et al.*, 2020] Jia-Chen Gu, Tianda Li, et al. Pre-trained and attention-based neural networks for building noetic task-oriented dialogue systems. In *DSTC8*, 2020.

[Gu *et al.*, 2021] Jia-Chen Gu, Chongyang Tao, Zhen-Hua Ling, et al. MPC-BERT: A pre-trained language model for multi-party conversation understanding. In *ACL*, 2021.

[Gu *et al.*, 2022] Jia-Chen Gu, Chao-Hong Tan, Chongyang Tao, Zhen-Hua Ling, et al. HeterMPC: A heterogeneous graph neural network for response generation in multi-party conversations. In *ACL*, 2022.

[Hawes *et al.*, 2009] Timothy Hawes, Jimmy Lin, and Philip Resnik. Elements of a computational model for multi-party discourse: The turn-taking behavior of supreme court justices. *JASIST*, 60(8):1607–1615, 2009.

[He *et al.*, 2013] Hua He, Denilson Barbosa, et al. Identification of speakers in novels. In *ACL*, 2013.

[Hu *et al.*, 2019] Wenpeng Hu, Zhangming Chan, Bing Liu, et al. GSN: A graph-structured network for multi-party dialogues. In *IJCAI*, 2019.

[Huang *et al.*, 2020] Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. Challenges in building intelligent open-domain dialog systems. *ACM TOIS*, 38(3):1–32, 2020.

[Iosif and Mishra, 2014] Elias Iosif and Taniya Mishra. From speaker identification to affective analysis: a multi-step system for analyzing children's stories. In *CLFL*, 2014.

[Jia *et al.*, 2020] Qi Jia, Yizhu Liu, Siyu Ren, Kenny Zhu, and Haifeng Tang. Multi-turn response selection using dialogue dependency relations. In *EMNLP*, 2020.

[Jiang *et al.*, 2018] Jyun-Yu Jiang, Francine Chen, Yan-Ying Chen, and Wei Wang. Learning to disentangle interleaved conversational threads with a siamese hierarchical network and similarity ranking. In *NAACL*, 2018.

[Jiang *et al.*, 2021] Ziyou Jiang, Lin Shi, Celia Chen, Jun Hu, and Qing Wang. Dialogue disentanglement in software engineering: How far are we? In *IJCAI*, 2021.

[Kim *et al.*, 2021] Seokhwan Kim, Michel Galley, Chulaka Gunasekara, et al. Overview of the eighth dialog system technology challenge: DSTC8. *IEEE/ACM TASLP*, 2021.

[Kummerfeld *et al.*, 2019] Jonathan K Kummerfeld, Sai R Gouravajhala, Joseph J Peper, et al. A large-scale corpus for conversation disentanglement. In *ACL*, 2019.

[Le *et al.*, 2019] Ran Le, Wenpeng Hu, et al. Who is speaking to whom? learning to identify utterance addressee in multi-party conversations. In *EMNLP*, 2019.

[Lee and Choi, 2021] Bongseok Lee and Yong Suk Choi. Graph based network with contextualized representations of turns in dialogue. In *EMNLP*, 2021.

[Li and Zhao, 2021] Yiyang Li and Hai Zhao. Self-and pseudo-self-supervised prediction of speaker and key-utterance for multi-party dialogue reading comprehension. In *Findings of EMNLP*, 2021.

[Li *et al.*, 2020a] Linxiao Li, Can Xu, et al. Zero-resource knowledge-grounded dialogue generation. In *NeurIPS*, 2020.

[Li *et al.*, 2020b] Tianda Li, Jia-Chen Gu, et al. DialBERT: A hierarchical pre-trained model for conversation disentanglement. 2020.

[Li *et al.*, 2021] Jiaqi Li, Ming Liu, et al. DADgraph: A discourse-aware dialogue graph neural network for multiparty dialogue machine reading comprehension. In *IJCNN*, 2021.

[Liu *et al.*, 2019] Cao Liu, Kang Liu, Shizhu He, et al. Incorporating interlocutor-aware context into response generation on multi-party chatbots. In *CoNLL*, 2019.

[Liu *et al.*, 2020] Hui Liu, Zhan Shi, Jia-Chen Gu, Quan Liu, Si Wei, and Xiaodan Zhu. End-to-end transition-based online dialogue disentanglement. In *IJCAI*, 2020.

[Liu *et al.*, 2021] Hui Liu, Zhan Shi, and Xiaodan Zhu. Unsupervised conversation disentanglement through co-training. In *EMNLP*, 2021.

[Lowe *et al.*, 2015] Ryan Lowe, Nissan Pow, Iulian Vlad Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*, 2015.

[Mahajan and Shaikh, 2021] Khyati Mahajan and Samira Shaikh. On the need for thoughtful data collection for multi-party dialogue: A survey of available corpora and collection methods. In *SIGDIAL*, 2021.

[Mehri and Carenini, 2017] Shikib Mehri and Giuseppe Carenini. Chat disentanglement: Identifying semantic reply relationships with random forests and recurrent neural networks. In *IJCNLP*, 2017.

[Meng *et al.*, 2017] Zhao Meng, Lili Mou, and Zhi Jin. Hierarchical RNN with static sentence-level attention for text-based speaker change detection. In *CIKM*, 2017.

[Meng *et al.*, 2018] Zhao Meng, Lili Mou, and Zhi Jin. Towards neural speaker modeling in multi-party conversation: The task, dataset, and models. In *LREC*, 2018.

[Ouchi and Tsuboi, 2016] Hiroki Ouchi and Yuta Tsuboi. Addressee and response selection for multi-party conversation. In *EMNLP*, 2016.

[O'Keefe *et al.*, 2012] Tim O'Keefe, Silvia Pareti, James R Curran, et al. A sequence labelling approach to quote attribution. In *EMNLP*, 2012.

[Pappadopulo *et al.*, 2021] Duccio Pappadopulo, Lisa Bauer, Marco Farina, et al. Disentangling online chats with dag-structured lstms. In **SEM*, 2021.

[Pareti *et al.*, 2013] Silvia Pareti, Tim O'keefe, Ioannis Konstas, et al. Automatically detecting and attributing indirect quotations. In *EMNLP*, 2013.

[Perret *et al.*, 2016] Jérémy Perret, Stergos Afantenos, Nicholas Asher, and Mathieu Morey. Integer linear programming for discourse parsing. In *NAACL*, 2016.

[Pinhanez *et al.*, 2018] Claudio S Pinhanez, Heloisa Candello, Mauro C Pichiliani, et al. Different but equal: Comparing user collaboration with digital personal assistants vs. teams of expert agents. 2018.

[Rajpurkar *et al.*, 2018] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In *ACL*, 2018.

[Riou *et al.*, 2015] Matthieu Riou, Soufian Salim, and Nicolas Hernandez. Using discursive information to disentangle french language chat. In *NLP4CMC*, 2015.

[Serban *et al.*, 2016] Iulian Serban, Alessandro Sordoni, Yoshua Bengio, et al. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, 2016.

[Shen *et al.*, 2006] Dou Shen, Qiang Yang, Jian-Tao Sun, and Zheng Chen. Thread detection in dynamic text message streams. In *SIGIR*, 2006.

[Shi and Huang, 2019] Zhouxing Shi and Minlie Huang. A deep sequential model for discourse parsing on multi-party dialogues. In *AAAI*, 2019.

[Tan *et al.*, 2019] Ming Tan, Dakuo Wang, Yupeng Gao, et al. Context-aware conversation thread detection in multi-party chat. In *EMNLP*, 2019.

[Tao *et al.*, 2019] Chongyang Tao, Wei Wu, Can Xu, et al. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *ACL*, 2019.

[Tao *et al.*, 2021] Chongyang Tao, Jiazhan Feng, Rui Yan, et al. A survey on response selection for retrieval-based dialogues. In *IJCAI*, 2021.

[Traum, 2004] David Traum. Issues in multiparty dialogues. In *Workshop on Agent Communication Languages*, 2004.

[Uthus and Aha, 2013] David C Uthus and David W Aha. Multiparticipant chat analysis: A survey. *Artificial Intelligence*, 2013.

[Vinyals *et al.*, 2015] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *NeurIPS*, 2015.

[Wang and Oard, 2009] Lidan Wang and Douglas W Oard. Context-based message expansion for disentanglement of interleaved text conversations. In *NAACL*, 2009.

[Wang *et al.*, 2020] Weishi Wang, Steven CH Hoi, and Shafiq Joty. Response selection for multi-party conversations with dynamic topic tracking. In *EMNLP*, 2020.

[Wang *et al.*, 2021] Ante Wang, Linfeng Song, Hui Jiang, et al. A structure self-aware model for discourse parsing on multi-party dialogues. In *IJCAI*, 2021.

[Wu *et al.*, 2017] Yu Wu, Wei Wu, et al. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *ACL*, 2017.

[Wu *et al.*, 2020] Shuangzhi Wu, Yufan Jiang, Xu Wang, et al. Enhancing response selection with advanced context modeling and post-training. In *AAAI*, 2020.

[Yu and Joty, 2020] Tao Yu and Shafiq Joty. Online conversation disentanglement with pointer networks. In *EMNLP*, 2020.

[Zhang *et al.*, 2018] Rui Zhang, Honglak Lee, et al. Addressee and response selection in multi-party conversations with speaker interaction rnns. In *AAAI*, 2018.

[Zhang *et al.*, 2020] Yizhe Zhang, Siqi Sun, Michel Galley, et al. DialoGPT: Large-scale generative pre-training for conversational response generation. In *ACL Demo*, 2020.

[Zhao *et al.*, 2019] Xueliang Zhao, Wei Wu, Chongyang Tao, et al. Low-resource knowledge-grounded dialogue generation. In *ICLR*, 2019.

[Zhou *et al.*, 2018] Xiangyang Zhou, Lu Li, Daxiang Dong, et al. Multi-turn response selection for chatbots with deep attention matching network. In *ACL*, 2018.

[Zhou *et al.*, 2020] Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. The design and implementation of xiaoice, an empathetic social chatbot. *CL*, 2020.

[Zhu *et al.*, 2020] Henghui Zhu, Feng Nan, et al. Who did they respond to? conversation structure modeling using masked hierarchical transformer. In *AAAI*, 2020.