Who Is Speaking to Whom? Learning to Identify Utterance Addressee in Multi-Party Conversations

Ran Le[13]*, **Wenpeng Hu**[2*], **Mingyue Shang**[1*], **Zhenjun You**[2*],
Lidong Bing [4], **Dongyan Zhao**[13] and **Rui Yan**[13†]
[1] Center for Data Science, AAIS, Peking University, Beijing, China
[2] School of Mathematical Sciences, Peking University, Beijing, China
[3] Wangxuan Institute of Computer Technology, Peking University, Beijing, China,
[4] R&D Center Singapore, Machine Intelligence Technology, Alibaba DAMO Academy
{leran.pku}@gmail.com {l.bing}@alibaba-inc.com
{wenpeng.hu,shangmy,youzhenjunpku,zhaody,ruiyan}@pku.edu.cn

Abstract

Previous research on dialogue systems generally focuses on the conversation between two participants, yet multi-party conversations which involve more than two participants within one session bring up a more complicated but realistic scenario. In real multi-party conversations, we can observe who is speaking, but the addressee information is not always explicit. In this paper, we aim to tackle the challenge of identifying all the missing addressees in a conversation session. To this end, we introduce a novel who-to-whom (**W2W**) model which models users and utterances in the session jointly in an interactive way. We conduct experiments on the benchmark Ubuntu Multi-Party Conversation Corpus and the experimental results demonstrate that our model outperforms baselines with consistent improvements.

1 Introduction

As an essential aspect of artificial intelligence, dialogue systems have attracted extensive attention in recent studies (Vinyals and Le, 2015; Serban et al., 2016). Researchers have paid great efforts to understand conversations between two participants, either single-turn (Li et al., 2016a; Shang et al., 2015; Vinyals and Le, 2015) or multi-turn (Zhou et al., 2016; Yan et al., 2016; Tao et al., 2019a,b), and achieved encouraging results. A more general and challenging scenario is that a conversation may involve more than two interlocutors conversing among each other (Uthus and Aha, 2013; Hu et al., 2019), which is known as multi-party conversation. Ubuntu Internet Relay Chat channel (IRC) is a multi-party conversation scenario as shown in Table 1. Generally, each utterance is associated with a speaker and one or more addressees in the conversation. Such a characteristic leads to complex speaker-addressee interactions. As a result, the speaker and addressee roles associated with utterances are constantly changing among multiple users across different turns. Such speaker and addressee information could be essential in many multi-party conversation scenarios including group meeting, debating and forum discussion. Therefore, compared to two-party conversations, a unique issue of multi-party conversations is to understand who is speaking to whom.

Table 1: An example of the multi-party conversation in the IRC dataset. Not all the addressees are specified.

Speaker	Utterance	Addressee
User 1	"Good point, tmux is the thing I miss."	–
User 1	"Cool thanks for ur help." @User 4	User 4
User 2	"Ahha, you r using something like cpanel."	–
User 3	"Yeah 1.4.0 exactly." @User 2	User 2
User 4	"my pleasure :)"	–

In real scenarios of multi-party conversations, an interesting phenomenon is that the speakers do not usually designate an addressee explicitly. This phenomenon also accords with our statistic analysis on the IRC dataset. We found that around 66% utterances missing explicit addressee information. That means when modeling such multi-party conversations, one may have to *guess* who is speaking to whom in order to understand the utterance correspondence as well as the stream structure of multi-party conversations.

Given a multi-party conversation where part of the addressees are unknown, previous work mainly focuses on predicting the addressee of only the last utterance. Ouchi and Tsuboi (2016) proposed to scan the conversation session and track the speaker's state based on the utterance content at each step. On this basis, Zhang et al. (2017) introduced a speaker interaction model that tracks all users' states according to their roles in the session. They both fused the representations of the last speaker and utterance as a query, and a match-

*Equal contribution.
†Corresponding author.

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 1909–1919, Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

ing network is utilized to calculate the matching degree between the query and each listener. The listener with the highest matching score is selected as the predicted addressee.

However, in practice, it is more helpful to predict all the missing addressees rather than only the last one in understanding the whole conversation. And it also benefits for both building a group-based chatbot and clustering users based on what they have said. Therefore, we propose a new task of identifying the addressees of all the missing utterances given a multi-party conversation session where part of the addressees are unspecified. To this end, we propose a novel Who-to-Whom (W2W) model which jointly models users and utterances in the multi-party conversation and predicts all the missing addressees in a uniform framework.[1] Our contributions are as follows:

- We introduce a new task of understanding who speaks to whom given an entire conversation session as well as a benchmark system.
- To capture the correlation within users and utterances in multi-party conversations, we propose an interactive representation learning approach to jointly learn the representations of users and utterances and enhance them mutually.
- The proposed approach (W2W) considers both previous and subsequent information in the session while incorporating the correlation with users and utterances. For conversations with complex structures, W2W models them in a uniform way and could handle any kind of occasion even when all the addressee information is missing.

2 Related Work

In this section, we briefly review recent works and progresses on multi-party conversations.

Multi-party conversations, as a general case of multi-turn conversations (Li et al., 2017, 2016c; Yan et al., 2016; Serban et al., 2016) involve more than two participants. In addition to the representation of learning for utterances, another key issue is to model multiple participants in the conversations. It is intuitive to introduce multiple user embeddings for multi-party conversations, either as persona-dependent embeddings (Li et al., 2016b), or as persona-independent embeddings (Ouchi and Tsuboi, 2016; Zhang et al., 2017; Meng et al., 2017). Recently, some researchers utilized users' information based on different roles in conversations, such as *senders* and *recipients* (Chen et al., 2017; Chi et al., 2017; Luan et al., 2016).

In multi-party conversations, identifying the relationship among users is also an important task. It can be categorized into two topics, 1) predicting who will be the next speaker (Meng et al., 2017) and 2) who is the addressee (Ouchi and Tsuboi, 2016; Zhang et al., 2017). For the first topic, Meng et al. (2017) investigated a temporal-based and a content-based method to jointly model the users and context. For the second topic, which is closely related to ours, Ouchi and Tsuboi (2016) proposed to predict the addressee and utterance given a context with all available information. Later, Zhang et al. (2017) proposed a speaker-interactive model, which takes users' role information into consideration and implements a role-sensitively state tracking process.

In our task, the addressee identification problem is quite different from (Ouchi and Tsuboi, 2016) and (Zhang et al., 2017). Both of their studies aimed to make predictions on whom the last speaker addresses to. While in this paper, we focus on the whole session and aim to identify all the missing addressees. By contrast, our task is a more challenging scenario since it relies on the correlation within all users and utterances to identify the speaker-addressee structure of the entire session.

3 Overview

3.1 Problem Formulation

Given an entire multi-party conversation S with length T, the sequence of utterances in it is defined as $\{u_t\}_{t=1}^{T}$. Each utterance is associated with a speaker a_t^{SPR} and an addressee a_t^{ADR}. a_t^{SPR} is observable across the entire session while a_t^{ADR} is mostly unspecified as shown in Table 1. Our task is to identify the addressees for all utterances within the conversation session. The predicted addressee is denoted as $\hat{a}_t^{ADR}$. Formally, we have following formulations:

$$\begin{aligned} QUERY &: \{(a_t^{\text{SPR}}, u_t)\}_{t=1}^{T} \\ PREDICTIONS &: \{\hat{a}_t^{\text{ADR}}\}_{t=1}^{T} \end{aligned} \tag{1}$$

Let $A(S)$ denote the user set in the session S, thus $A(S)\backslash\{a_t^{SPR}\}$ denotes the listeners at the t-th turn ($a_t^{LSR_j}$ denotes the j-th listener). The listeners are also referred as candidate addressees for each turn and the identified addressee $\hat{a}_t^{ADR}$ should be one

[1] To make the model practical in learning, we assume that one utterance is associated with only one addressee.

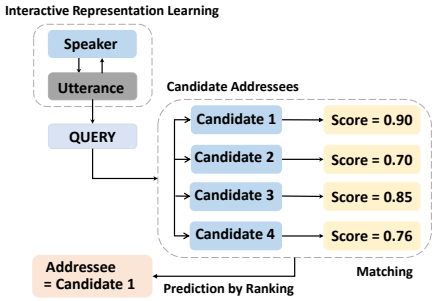

Figure 1: The system architecture of W2W model.

of them.[2]

3.2 Architecture Overview of W2W Model

Figure 1 illustrates the overview picture of the proposed W2W model which consists of a representation learning module and a matching module.

Concretely, the representation learning module is designed to jointly learn the representation of users and utterances in an interactive way after initializing them separately. The representations of users (also denoted as user states) and utterance embeddings are mutually enhanced. With the representations of users and utterances, a network is utilized to fuse them up into a query representation. In this way, we jointly capture who is speaking what at each step.

After the representations of users and utterances are learned, we feed them into a matching module. In this module, a matching network is learned to score the matching degrees between the query and each candidate. According to the matching scores, the model ranks all addressee candidates in $A(S)\backslash\{a_t^{SPR}\}$ and selects the one with the highest matching score as the identified addressee. For each utterance in the multi-party conversation, we repeat the above steps until the addressees of all utterances are identified.

4 Our W2W Model

In this section, we first describe each part of the W2W model in details: (1) Initialization of utterance and user representations; (2) Interactive representation learning of users and utterances; (3) Matching procedure for identifying the addressee. We finally describe the training procedure of the W2W model.

[2]In this paper, we denote vectors with bold lower-case (like $\mathbf{u}_i$) and matrices with bold upper-case like ($\boldsymbol{U}$ and $\boldsymbol{W}$).

4.1 Initialization

W2W models utterance and user embeddings separately before interactive representation learning and gets the representation of each utterance and user as initialization.

4.1.1 Utterance Initialization Encoder

Suppose that in a conversation session S with T utterances denoted as $\{u_1, u_2, \ldots, u_T\}$. An utterance u_t that contains n tokens is denoted as $\{w_1, w_2, \ldots, w_n\}$, where $\{\boldsymbol{w}_i\}$ is word embeddings[3] of the i-th token. We first utilize a word level bi-directional RNN with Gated Recurrent Units (GRUs) (Cho et al., 2014) to encode each utterance and take the concatenation of the hidden states of the last step from both sides as the sentence embedding. Then, a sentence level bi-directional GRU is applied with each sentence embedding as input to obtain the global context of the session. The utterance representation $\boldsymbol{u}_t$ is represented by the concatenation of hidden states from both sides at t-th time step. [4]

4.1.2 Position-Based User Initialization

In multi-party conversation, position information of different participants in the session is crucial in the addressee identification task. For example, a speaker is more likely to address his direct preceding or subsequent speaker. On this basis, we define the initialization user matrix $\mathbf{A}_{(0)}$ based on the speaking order of users in the session (Ouchi and Tsuboi, 2016). Concretely, all users in a session are sorted in a descending order according to the first time when they speak, and the i-th user is assigned with the i-th row of $\mathbf{A}_{(0)}$ as $\mathbf{a}^{\mathrm{i}}_{(0)}$. The user matrix $\mathbf{A}_{(0)}$ is trained as parameters along with other weight matrices in the neural network.

Users of the same order in different sessions share the same initialization user embedding. Note that the user representations are independent of each personality (unique user). Such strategy guarantees the initialization user embeddings to carry position information as well as handle new users unseen in training data during addressee identification.

[3] We use GloVe(Pennington et al., 2014), but it can be any word embeddings(Mikolov et al., 2013; Hu et al., 2016).

[4]Such a hierarchical framework (Serban et al., 2016) takes into account the context of all previous and future sentences in the whole session, thus enables the model to learn a strong representation.

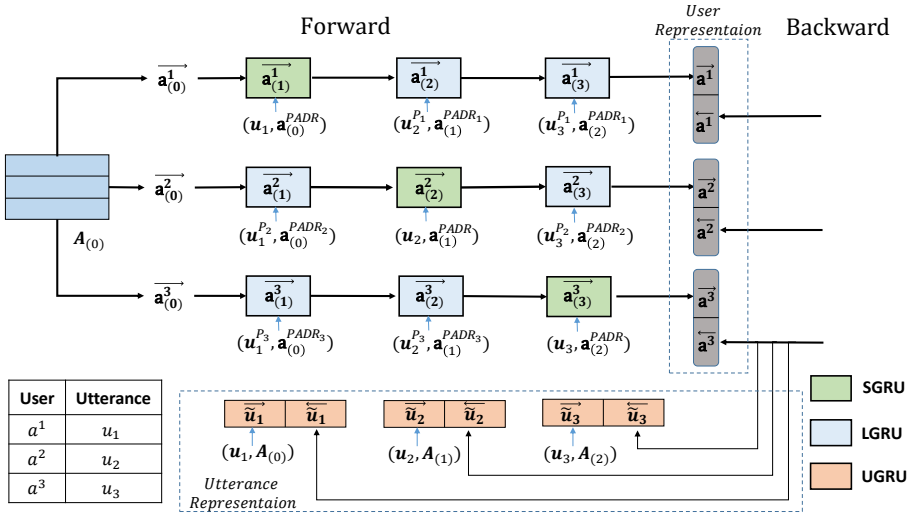

Figure 2: Interactive representation learning in W2W: At each utterance turn, SGRU tracks the speaker's state, LGRU tracks listeners' states and UGRU fuses users' states into utterances. W2W scans the conversation session from two directions and the representation of each user and utterance is the concatenation of both sides.

4.2 Interactive Representation Learning

To better capture who speaks what at each turn through the whole session, we propose to interactively learn the representation of utterances and users. Different from prior studies (Ouchi and Tsuboi, 2016; Zhang et al., 2017) which only track the users' states but neglecting the users' impact on utterances. We propose the W2W model which learns user and utterance representations interactively by tracking users' states with utterance embeddings as well as fusing users' states into the utterance embeddings.

4.2.1 Users Representation Learning

Role-sensitive User State Tracking. Suggested by (Zhang et al., 2017), an utterance could have different degrees of impact on the states of the corresponding speaker and listeners. In order to capture the users' role information, we utilize two kinds of GRU-based cells represented as Speaker-GRU (SGRU) and Listener-GRU (LGRU) to track the states of the speaker and listeners respectively at each turn of the session.[5] At the t-th transition step, the SGRU tracks the speaker representation $\mathbf{a}^{\mathrm{SPR}}_{(\mathrm{t})}$, from the former state of him $\mathbf{a}^{\mathrm{SPR}}_{(\mathrm{t}-1)}$, the utterance representation $\mathbf{u}_t$, as well as a pseudo addressee representation $\mathbf{a}^{\mathrm{PADR}}_{(\mathrm{t}-1)}$ calculated via PAM (Person Attention Mechanism) which is a weighted sum of all the listeners' representations. Details on PAM will be elaborated in the next part. The state tracking procedure for the i-th step is formulated as Eq (2). The main idea of SGRU is to incorporate two reset gates, each of which controls the information fusion from the listeners and speaker respectively, denoted as r_i and p_i. $\boldsymbol{W}$, $\boldsymbol{U}$ and $\boldsymbol{V}$ are learnable parameters.

$$
\begin{aligned}
\mathbf{r}_i &= \sigma(\boldsymbol{W}_r\boldsymbol{u}_i + \boldsymbol{U}_r\mathbf{a}^{\mathrm{SPR}}_{(\mathrm{i}-1)} + \mathrm{V}_{\mathrm{r}}\mathbf{a}^{\mathrm{PADR}}_{(\mathrm{i}-1)}) \\
\mathbf{p}_i &= \sigma(\boldsymbol{W}_p\boldsymbol{u}_i + \boldsymbol{U}_p\mathbf{a}^{\mathrm{SPR}}_{(\mathrm{i}-1)} + \mathrm{V}_{\mathrm{p}}\mathbf{a}^{\mathrm{PADR}}_{(\mathrm{i}-1)}) \\
\mathbf{z}_i &= \sigma(\boldsymbol{W}_z\boldsymbol{u}_i + \boldsymbol{U}_z\mathbf{a}^{\mathrm{SPR}}_{(\mathrm{i}-1)} + \mathrm{V}_{\mathrm{z}}\mathbf{a}^{\mathrm{PADR}}_{(\mathrm{i}-1)}) \\
\widetilde{\mathbf{a}}^{SPR}_{(i)} &= \tanh(\boldsymbol{W}\boldsymbol{u}_i + \boldsymbol{U}(\mathbf{r}_i \odot \mathbf{a}^{\mathrm{SPR}}_{(\mathrm{i}-1)}) + \mathrm{V}(\mathbf{p}_{\mathrm{i}} \odot \mathbf{a}^{\mathrm{PADR}}_{(\mathrm{i}-1)})) \\
\mathbf{a}^{\mathrm{SPR}}_{(\mathrm{i})} &= \mathbf{z}_i \odot \mathbf{a}^{\mathrm{SPR}}_{(\mathrm{i}-1)} + (1-\mathbf{z}_{\mathrm{i}}) \odot \widetilde{\mathbf{a}}^{\mathrm{SPR}}_{(\mathrm{i})}
\end{aligned}
\tag{2}
$$

Symmetrically, LGRU incorporates the embeddings of a certain listener as well as a pseudo speaker and a pseudo utterance representation (also calculated via PAM) as inputs and tracks the state of each listener. SGRU and LGRU have symmetric updating functions as Eq (2) except for the difference on pseudo representation incorporated in the cell.[6] The parameters of SGRU and LGRU are not shared, which guarantees W2W to learn role-dependent features in users' state tracking procedure. The whole structure of SGRU and LGRU are illustrated in Figure 3.

Person Attention Mechanism. We propose a person attention mechanism (PAM) (Eq (3) to model

[5] We denote a user embedding tracked until t_{th} time step as $\mathbf{a}_{(\mathrm{t})}$, with $\mathbf{a}^{\mathrm{SPR}}_{(\mathrm{t})}$ as the representation of the speaker at t_{th} turn and $\mathbf{a}^{\mathrm{LSR_j}}_{(\mathrm{t})}$ as the representation of the j_{th} listener at t_{th} turn.

[6] SGRU incorporates pseudo addressee $\mathbf{a}^{\mathrm{PADR}}_{(\mathrm{i}-1)}$ for speaker state tracking. LGRU incorporates the pseudo speaker $\mathbf{a}^{\mathrm{PSPR_j}}_{(\mathrm{i}-1)}$ and pseudo utterance $\mathbf{u}^j_i$ to track each j_{th} listener

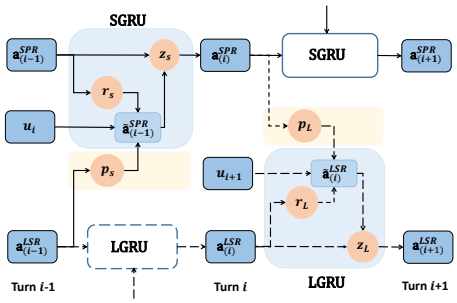

Figure 3: Structures of SGRU and LGRU.

the state tracking process exactly. Each element β_i^j measures how likely the model estimates the j-th listener to be the addressee for the i-th turn based on the user representations tracked until turn i. $\boldsymbol{W}_p$ is the parameter.

$$\beta_i^j = \sigma(\mathbf{a}^{\mathrm{LSR_j}}_{(i-1)} \mathbf{W}_{\mathrm{p}} [\mathbf{a}^{\mathrm{SPR}}_{(i-1)}; \mathbf{u}_i]^{\mathsf{T}}) \tag{3}$$

Then for each turn i, a pseudo addressee $\mathbf{a}^{\mathrm{PADR}}_{(i-1)}$ is generated as the weighted sum of all listener representations tracked until step i as Eq (4). Intuitively, a listener with a higher matching score is more likely to be the addressee at the current step. The pseudo addressee $\mathbf{a}^{\mathrm{PADR}}_{(i-1)}$ is incorporated into the state tracking of the speaker as Eq (2).

$$\mathbf{a}^{\mathrm{PADR}}_{(i-1)} = \frac{\sum_j \beta_i^j \cdot \mathbf{a}^{\mathrm{LSR_j}}_{(i-1)}}{\sum_j \beta_i^j} \tag{4}$$

$$\mathbf{a}^{\mathrm{PSPR_j}}_{(i-1)} = \beta_i^j \cdot \mathbf{a}^{\mathrm{SPR_j}}_{(i-1)} \tag{5}$$

$$\mathbf{u}_i^{P_j} = \beta_i^j \cdot \mathbf{u}_i \tag{6}$$

Symmetrically, the pseudo speaker $\mathbf{a}^{\mathrm{PSPR_j}}_{(i-1)}$ and pseudo utterance $\mathbf{u}_i^{P_j}$ are generated through Eq (5) and Eq (6) for each listener j at the i-th turn of the conversation.

4.2.2 Utterances Representation Learning

We design a UGRU (Utterance-GRU) cell[7], which has the same structure as SGRU/LGRU to fuse the utterance embedding, current speaker embedding and the user-summary vector into an enhanced utterance embedding. The user matrix initialized with $\boldsymbol{A}_{(0)}$ (as described in 4.2.1) and tracked until step $t-1$ is denoted as $\boldsymbol{A}_{(t-1)}$. The user-summary vector is calculated through max-pooling over users on $\boldsymbol{A}_{(t-1)}$ as a summary of all users' current states.

$$\begin{aligned} \boldsymbol{h}_s &= \text{Max-Pool}(\boldsymbol{A}_{(t-1)}) \\ \tilde{\boldsymbol{u}}_t &= \text{UGRU}(\boldsymbol{u}_t, \mathbf{a}^{\mathrm{SPR}}_{(t-1)}, \mathrm{h_s}) \end{aligned} \tag{7}$$

[7]Note that although UGRU has the same structure as SGRU/LGRU, it acts on each utterance for only once instead of recurrently tracking.

Input: Initial representations of utterances $\{\boldsymbol{u}_t\}_{t=1}^{T}$
Initial user matrix $\boldsymbol{A}_{(0)}$
for $i = 0; i < T; i++$ **do**
  Calculate current matching scores through **PAM** Eq (3);
  Generate pseudo embeddings using Eq (4,5,6);
  Track the speaker state through SGRU using Eq (2);
  Track each listener's state with LGRU using Eq (2);
  Fuse users' information into utterance representation using Eq (7) ;
end
return *1) User matrix of the last turn* $\overrightarrow{\boldsymbol{A}}_{(T)}$;
2) Enhanced utterance embeddings $\{\overrightarrow{\tilde{\boldsymbol{u}}}_t\}_{t=1}^{T}$.

Algorithm 1: Interactive Representation Learning Algorithm (Forward Pass).

4.2.3 Forward-and-Backward Scanning.

Considering that the addressee of an utterance can be the speaker of the preceding utterances or the subsequent ones, it is important to capture the dependency from both sides for users and utterances. We propose a forward-and-backward scanning schema to enhance the interactive representation learning. For forward pass, W2W model outputs the forward user matrix of the last time step, denoted as $\overrightarrow{\boldsymbol{A}}_{(T)}$ as well as all the forward-enhanced utterance embeddings $\{\overrightarrow{\tilde{\boldsymbol{u}}}_t\}_{t=1}^{T}$ as illustrated in Algorithm 1. The backward pass initializes users and utterances in the same way as the forward pass and scans the conversation session in the reversed order. Representations from both sides are concatenated correspondingly as the final representation as Eq (8).

$$\tilde{\mathbf{u}}_i = [\overrightarrow{\tilde{\boldsymbol{u}}_i}; \overleftarrow{\tilde{\boldsymbol{u}}_i}]; \quad \mathbf{a}^{\mathrm{j}} = [\overrightarrow{\mathbf{a}^{\mathrm{j}}_{(\mathrm{T})}}; \overleftarrow{\mathbf{a}^{\mathrm{j}}_{(\mathrm{T})}}] \tag{8}$$

4.3 Matching

Matching Network. We first fuse the speaker embedding and the utterance embedding into a query representation as $\mathbf{q}$, then measures the embedding similarity s between the query and each listener:

$$\begin{aligned} \mathbf{q} &= \tanh(\mathbf{W}_s \mathbf{a}^{\mathrm{SPR}} + \mathbf{W}_{\mathrm{u}} \tilde{\mathbf{u}}) \\ s &= \sigma(\mathbf{a}^{\mathrm{LSR}} \mathbf{W}_{\mathrm{m}} \mathbf{q}^{\mathsf{T}}) \end{aligned} \tag{9}$$

Table 2: Data statistics: sample size of datasets with different session lengths (i.e., 5, 10, 15) from the Ubuntu dataset (Ouchi and Tsuboi, 2016).

Dataset	Train	Dev	Test
Len-5	461,120	28,570	32,668
Len-10	495,226	30,974	35,638
Len-15	489,812	30,815	35,385

where $\boldsymbol{W}_s, \boldsymbol{W}_u, \boldsymbol{W}_m$ denote weight matrices. For simplicity, we use a short-handed $Match(.)$ to denote the Equation (9) when there is no ambiguity.

Addressee Identification. For each turn in the session, we score the matching degree between the query and each listener and select the best matched $\hat{a}_i^{ADR}$ as the addressee prediction as Eq (10). s_i^j denotes the matching score between the j-th listener $\mathbf{a}^{\mathrm{LSR_j}}$ and the query of the i-th turn. For the entire conversation, we repeat the above steps until the addressee for each utterance is identified.

$$\begin{aligned} s_i^j &= \mathrm{Match}(\mathbf{a}_i^{\mathrm{SPR}}, \mathbf{a}_i^{\mathrm{LSR_j}}, \widetilde{\mathbf{u}}_i) \\ \hat{a}_i^{ADR} &= \mathrm{argmax}_j(\{s_i^j\}) \end{aligned} \tag{10}$$

4.4 Learning

We utilize the cross-entropy loss to train our model (Ouchi and Tsuboi, 2016). The objective is to minimize the loss as follows:

$$loss = -\sum_k \sum_i \left(\log(s_i^+) + \log\left(1 - s_i^-\right)\right) + \frac{\lambda}{2}\|\theta\|_2 \tag{11}$$

Each subscript k denotes a session, and subscript i is taken from the utterances that have ground truths of addressee information. s_i^+ denotes the matching score between the query and the ground truth addressee, s_i^- denotes the score of the negative matching, where the candidate addressee is negatively sampled. All parameters in our W2W model are jointly trained via Back-Propagation (BP) (Rumelhart et al., 1986).

5 Experimental Setups

Dataset. We run experiments using the benchmark Ubuntu dataset released by Ouchi and Tsuboi (2016). The corpus consists of a huge amount of records including response utterances, user IDs and their posting time. We organize the dataset as samples of conversation sessions.

We filter out the sessions without a single addressee ground truth, which means no label is available in these sessions. We also filter out session samples with one or more blank utterance. Moreover, we separate the conversation sessions into three categories according to the session length. Len-5 indicates the sessions with 5 turns and it is similar for Len-10 and Len-15. Such a splitting strategy is adopted in related studies as (Ouchi and Tsuboi, 2016; Zhang et al., 2017). The dataset is split into train-dev-test sets and the statistics are shown in Table 2.

Comparison Methods. We utilize several algorithms including heuristic and state-of-the-art methods as baselines. As there is no existing method that can perform the new task, we have to adapt baselines below into our scenario.

• Preceding: The addressee is designated as the preceding speaker of the current speaker.

• Subsequent: The addressee is designated as the next speaker.

• Dynamic RNN (DRNN): The model is originally designed to predict the addressee of the last utterance given the whole context available (Ouchi and Tsuboi, 2016). We adapt it to our scenario which is to identify addressees for all utterances in the conversation session. Concretely, the representations of users and context are learned in the same way as DRNN. While during the matching procedure, the representations of speaker and context are utilized to calculate the matching degree with candidate addressees at each turn.

• Speaker Interactive RNN (SIRNN): SIRNN is an extension model on DRNN, which is more interaction-sensitive (Zhang et al., 2017). Since all addressee information is totally unknown in our scenario, we also adapt the model with only speaker-role and observer-role into this situation. User states are tracked recurrently according to their roles at each turn, i.e. two distinct networks ($IGRU^S$ and $IGRU^O$) are utilized to track the status of the speaker and observers at each turn. Since there is no addressee-role observable through the session, we also make some adaption on the updating cell here. At each turn, $IGRU^S$ updates the speaker embedding from the previous speaker embedding and the utterance embedding, $IGRU^O$ updates the observer embedding from the previous observer embedding and the utterance embedding. During matching procedure, we make the prediction on each turn instead of only predicting the addressee for the last turn.

Implementation and Parameters. For fair comparison, we choose the hyper-parameters spec-

Table 3: Addressee identification results (in %) on datasets (Len-5, Len-10 and Len-15) where $\star$ denotes p-value < 0.01 in the significance test against all the baselines.

	Len-5				Len-10				Len-15			
Model	p@1	p@2	p@3	Acc.	p@1	p@2	p@3	Acc.	p@1	p@2	p@3	Acc.
Preceding	63.50	90.05	98.83	40.46	56.84	80.15	91.86	21.06	54.97	77.19	88.75	13.08
Subsequent	61.03	88.86	98.54	40.25	54.57	73.60	87.26	20.26	53.07	69.85	81.93	12.79
DRNN	72.75	93.21	99.24	58.18	65.58	85.85	94.92	34.47	62.60	82.68	92.14	22.58
SIRNN	75.98	94.49	99.39	62.06	70.88	89.14	96.10	40.66	68.13	85.82	93.52	28.05
W2W	**77.55***	**95.11***	**99.57**	**63.81***	**73.52***	**90.33***	**96.64**	**44.14***	**73.42***	**89.44***	**95.51***	**34.23***

Table 4: Consistency comparison between human inference and model predictions on overlapping rate (%). $\star$ denotes p-value < 0.01 in the significance test against all the baselines.

Model	Len-5	Len-10	Len-15
DRNN	75.60	67.54	63.06
SIRNN	78.94	71.59	67.22
W2W	**80.86***	**74.05***	**71.14***

ified by Ouchi and Tsuboi (2016) and Zhang et al. (2017). We represent the words with 300-dimensional GloVe vectors, which are fixed during training. The dimension of speaker embeddings and hidden states are set to 50. The joint cross-entropy loss function with L_2 weight decay as 0.001 is minimized by Adam (Kingma and Ba, 2014) with a batch size of 128.

Evaluation Metrics. To examine the effectiveness of our model on the addressee identification task, we compare it with baselines in terms of precision@n (i.e. p@n) (Yan et al., 2017). For predicting an addressee of an utterance, our model actually provides a ranking list for all candidate addressees.[8] We also evaluate the performance on the session level: we mark a session as a positive sample if and only if all ground truth labels are correctly identified on the top 1 of rankings, and calculate the ratio as accuracy.

As we discussed before, only a part of utterances in multi-party conversations have explicit addressee, which limits the completeness of automatic evaluation metrics. In order to evaluate the performance on unlabeled utterances, we leverage human inference results and calculate the consistency between the model predictions and human results. Due to the labor cost limit, we randomly sample 100 sessions from the test set of Len-5, Len-10 and Len-15 respectively and recruit three volunteers to annotate the addressee label for unlabeled utterances by reasoning through content and addressee information. We leave blank on the utterance where three annotators give different inference results. Finally around 81.4% of the unlabeled utterances have two or more annotators given them same annotation. With the human inference results and model predictions, we use the overlapping rate[9] as the consistency metric.

6 Results and Discussion

We first give an overall comparison between W2W and baselines followed by ablation experiments to demonstrate the effectiveness of each part in W2W model. We then confirm the robustness of W2W with several factors including the numbers of users, the position of the utterance. Furthermore, we evaluate how W2W and baseline models perform on both labeled and unlabeled utterance.

Overall Performance. For automatic evaluation shown in Table 3, end-to-end deep learning approaches outperform heuristic ones, which indicates that simple intuition is far from satisfaction and further confirm the value of this work. Among the deep-learning based approaches, our W2W model outperforms the state-of-the-art models by all evaluation metrics. Direct adaption from approaches on identifying the last addressee of the session may not work fine for our scenario.

As shown in Figure 4, the performance of all methods drops as the context length increases (from Len-5 to Len-15) since the task becomes more difficult with more context information to encode and more candidate addressees to rank. However, the improvement of our W2W model is rather more obvious with longer context length. In particular, for the dataset Len-15, W2W improves 5% on p@1 and 10% on session accuracy over SIRNN as shown in Figure 4, which indicate the robustness of W2W model in complex scenarios.

[8] Intuitively, p@1 is the precision at the highest ranked position and should be the most natural way to indicate the performance. We also provide p@2 and p@3 to illustrate the potential of different systems to identify the correct addressee on top of the lists.

[9] The calculation formula of the overlapping rate is described in Appendix.

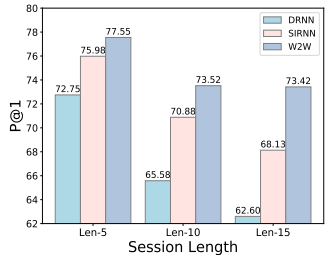

Figure 4: The comparison between W2W model and two state-of-the-art baselines on p@1.

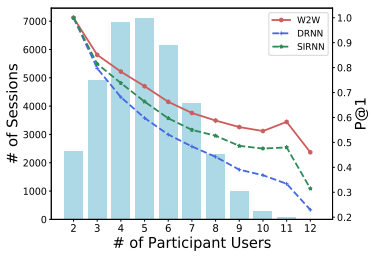

Figure 5: p@1 performance vs. number (#) of participants within a session. Results are tested on Len-15.

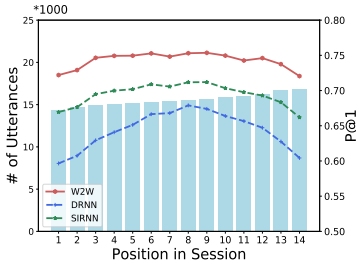

Figure 6: p@1 performance vs. positions of addressees within a session. Results are tested on Len-15.

Table 4 shows the consistency between human inference and model predictions. W2W also outperforms the baselines with a larger margin on longer conversation scenarios, which is consistent with the phenomenon of automatic evaluation. The advantage on unlabeled data of our W2W model demonstrates the superiority for detecting the latent speaker-addressee structure unspecified in the conversation stream, and that it could help find out the relationship between and across users in the session.

Ablation Test. Table 5 shows the results of ablation test. First, we replace the bi-directional scanning schema with the forward scanning one. The result shows that the bi-directional scanning schema captures the information in long conversations more sufficiently. Besides, we investigate the effectiveness of PAM by replacing it with the simple mean-pooling approach as Eq (12):

$$\begin{aligned} \mathbf{a}^{\mathrm{PADR}}_{(\mathrm{i}-1)} &= \frac{1}{n} \cdot \sum_{j} \mathbf{a}^{\mathrm{LSR_j}}_{(\mathrm{i}-1)} \\ \mathbf{a}^{\mathrm{PSPR_j}}_{(\mathrm{i}-1)} &= \frac{1}{n} \cdot \mathbf{a}^{\mathrm{SPR_j}}_{(\mathrm{i}-1)} \\ \mathbf{u}_i^{P_j} &= \frac{1}{n} \cdot \mathbf{u}_i \end{aligned} \tag{12}$$

The result shows that it is more beneficial to capture the correlation between the user-utterance pair and each listener and implement the state tracking correspondingly at each turn with our PAM mechanism.

To investigate the effectiveness of interactive representation learning module, we first remove the UGRU cell and fix the utterance representations in the state tracking procedure (referred as w/o Utterance Interaction in Table 5). Symmetrically, we fix the user representations in the session by removing the SGRU and LGRU cell and maintain only the interaction affect from the users to the utterances (referred as w/o User Interaction in Table 5). We also conduct an experiment on taking off the whole interactive representation module by removing UGRU, SGRU and LGRU, where the addressee identification is totally dependent on the initial representations of users and utterances. The result demonstrates that each part of them has an important contribution to our W2W model, especially the interaction affect on users.

Table 5: Ablation test on the effectiveness of W2W model parts in dataset Len-15.

Model	$p@1$	$p@2$	$p@3$	Acc.
W2W w/ Forward Scanning Only	71.60	87.99	94.80	31.39
W2W w/o PAM	72.56	88.83	95.21	32.78
W2W w/o Utterance Interaction	72.94	88.89	95.28	33.04
W2W w/o User Interaction	49.18	72.38	86.81	18.24
W2W w/o Interaction	46.39	71.66	86.14	15.15
W2W	**73.42**	**89.41**	**95.50**	**33.89**

Number of Participants. The task becomes more difficult as the number of participants involved in the conversation increases since more participants correspond to more complicated speaker-addressee relationship. We investigate how the performance correlates with the number of speakers in the dataset Len-15. The results in terms of p@1 are shown in Figure 5. In conversations with few participants, all methods have rather high performance. As the participant number increases, W2W constantly outperforms the baselines. The performance gap becomes larger especially when there are 6 users and more, which indicates the capability of our W2W model in handling complex conversation situations.

Position of Addressee-to-Predict. As mentioned above in 4.1.2, position information of utterances is a crucial factor in identifying addressees for multi-party conversation. We investigate how the system performance correlates with the position of the addressee to be predicted. In

Figure 6, we show the p@1 performance of the W2W and baselines when predicting the addressee of u_i at the i_{th} turn. Again, W2W shows consistently better performance than the other baselines no matter where the turn to predict addressee is.

We can observe that all the methods perform relatively poor at the beginning or the end. Compared with the middle part of a long conversation, the beginning and the ending contains less context information, which makes the addressee of these part more difficult to predict. The result in Figure 6 shows that the gap between W2W and other methods is even larger where the addressee-to-predict is at the beginning or the end, which indicates that W2W is better at capturing key information and has stronger robustness in difficult scenarios.

Variance of Matching Scores. In real multi-party conversations, the utterances without addressee information can be divided into two cases. Sometimes an utterance has an explicit addressee while the speaker doesn't specify whom he/she is speaking to. We denote these cases as **NP** which refers to NULL-Positive. Another case is that the utterances don't address to any user in the conversation (denoted as **NN** which means Null-Negative), such as utterances 'Hi, everyone!' and 'Can anyone here help me?' In Ubuntu IRC dataset, unlabeled of the NN case and NP case are mixed and are difficult to distinguish without manual annotation.

Meanwhile, our W2W model and baseline approaches predict matching scores on each listener for every utterance no matter it has addressee information or not. For each utterance, the variance of the matching scores on all listeners represents how certain the model is on its addressee identification decision. A larger variance corresponds to a more confident prediction. Table 6 demonstrates the variance comparison on labeled and unlabeled cases in the test set Len-15. On utterances with addressee labels, the variance of our W2W model is significantly larger than the state-of-the-art baseline, which indicates that W2W has a higher degree of certainty about its own predictions when the conversation content is referring to someone explicitly. For utterances without addressee labels, the difference of variance between W2W and SIRNN is significantly reduced. Considering that unlabeled sets consist of NN ones as well as NP ones on which W2W has much larger variance than SIRNN just as the labled utterance scenario, we can infer that W2W has much lower variance than SIRNN on the NN case. Such phenomenon reflects that our W2W model won't make a prediction recklessly on occasions where there is no clear addressee. Therefore, the variance on matching scores of all listeners in our W2W model, to some extent, provides a signal of whether the utterance has a explicit addressee even if we do not provide any supervision information on this aspect during training.

Table 6: Variance of matching scores on labeled and unlabeled utterances in the set Len-15.

Model	Labeled	Unlabeled
W2W	0.111	0.080
SIRNN	0.088	0.077

7 Conclusion

In this paper, we aim at learning to identify the utterance addressees in multi-party conversations by predicting who is speaking to whom, which is a new task.. To perform the new task, we propose the W2W model which learns the user and utterance representations interactively. To handle the uncertainty in the conversation session, we design PAM which captures matching degree between current query and each candidate addressees. The experimental results show prominent and consistent improvement over heuristic and state-of-the-art baselines.

In the future, we will further investigate better utterance models associated with additional information such as topics or knowledge. With the help of learned addressee structure, we can build a general structural dialogue system for complex multi-party conversations.

Acknowledgments

We would like to thank the reviewers for their constructive comments. We would also like to thank Bing Liu and Zhangming Chan from Peking University for their suggestion and help on this paper. This work was supported by the National Key Research and Development Program of China (No. 2017YFC0804001), the National Science Foundation of China (NSFC No. 61876196 and NSFC No. 61672058). Rui Yan and Wenpeng Hu were sponsored by Alibaba Innovative Research (AIR) Grant.

References

Po-Chun Chen, Ta-Chung Chi, Shang-Yu Su, and Yun-Nung Chen. 2017. Dynamic time-aware attention to speaker roles and contexts for spoken language understanding. In *Automatic Speech Recognition and Understanding Workshop (ASRU), 2017 IEEE*, pages 554–560, Okinawa Japan. IEEE.

Ta Chung Chi, Po Chun Chen, Shang-Yu Su, and Yun-Nung Chen. 2017. Speaker role contextual modeling for language understanding and dialogue policy learning. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, volume 2, pages 163–168, Taipei Taiwan. IJCNLP.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha Qatar. EMNLP.

Wenpeng Hu, Zhangming Chan, Bing Liu, Dongyan Zhao, Jinwen Ma, and Rui Yan. 2019. Gsn: A graph-structured network for multi-party dialogues. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5010–5016. International Joint Conferences on Artificial Intelligence Organization.

Wenpeng Hu, Jiajun Zhang, and Nan Zheng. 2016. Different contexts lead to different word embeddings. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 762–771, Osaka, Japan. The COLING 2016 Organizing Committee.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 9:15.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, Berlin Germany. ACL.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 994–1003, Berlin Germany. ACL.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016c. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Berlin Germany. ACL.

Jiwei Li, Will Monroe, Tianlin Shi, Sèbastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169, Copenhagen Denmark. EMNLP.

Yi Luan, Yangfeng Ji, and Mari Ostendorf. 2016. Lstm based conversation models. *arXiv preprint arXiv:1603.09457*, 1.

Zhao Meng, Lili Mou, and Zhi Jin. 2017. Towards neural speaker modeling in multi-party conversation: The task, dataset, and models. *arXiv preprint arXiv:1708.03152*, 1.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2133–2143, Austin Texas USA. EMNLP.

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

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323(6088):533.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 3776–3783, Phoenix Arizona USA. AAAI Press, AAAI.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, volume 1, pages 1577–1586, Beijing China. ACL.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019a. Multi-representation fusion network for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, WSDM '19, pages 267–275, New York, NY, USA. ACM.

Chongyang Tao, Wei Wu, Can Xu, Wenpeng Hu, Dongyan Zhao, and Rui Yan. 2019b. One time of interaction may not be enough: Go deep with an interaction-over-interaction network for response selection in dialogues. In *Proceedings of the 57th*

Conference of the Association for Computational Linguistics, pages 1–11.

David C Uthus and David W Aha. 2013. The ubuntu chat corpus for multiparticipant chat analysis. In *Proceedings of the 27th AAAI Conference on Artificial Intelligence*, Bellevue Washington USA. AAAI.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*, 1.

Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 55–64, Pisa Tuscany Italy. ACM, SIGIR.

Rui Yan, Dongyan Zhao, et al. 2017. Joint learning of response ranking and next utterance suggestion in human-computer conversation system. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 685–694, Tokyo Japan. ACM, SIGIR.

Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. 2017. Addressee and response selection in multi-party conversations with speaker interaction rnns. *arXiv preprint arXiv:1709.04005*, 1.

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