Dialogue Structure Parsing on Multi-Floor Dialogue Based on Multi-Task Learning

Seiya Kawano[1] , **Koichiro Yoshino**[1,2] , **David Traum**[3] , **Satoshi Nakamura**[1,2]
[1]Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara, Japan.
[2]RIKEN, Center for Advanced Intelligence Project AIP, Tokyo, Japan.
[3]USC Institute for Creative Technologies, Los Angeles, CA, USA.
{kawano.seiya.kj0, koichiro}@is.naist.jp, traum@ict.usc.edu, s-nakamura@is.naist.jp

Abstract

A multi-floor dialogue consists of multiple sets of dialogue participants, each conversing within their own floor, but also at least one multi-communicating member who is a participant of multiple floors and coordinating each to achieve a shared dialogue goal. The structure of such dialogues can be complex, involving intentional structure and relations that are within or across floors. In this study, we propose a neural dialogue structure parser based on multi-task learning and an attention mechanism on multi-floor dialogues in a collaborative robot navigation domain. Our experimental results show that our proposed model improved the dialogue structure parsing performance more than those of single models, which are trained on each dialogue structure parsing task in multi-floor dialogues.

1 Introduction

In single-floor dialogues, each participant can access any of the dialogue contents. For example, two people talking face to face, or an online conference involving participants from different places is a single-floor dialogue because each participant can access all of the dialogue contents. By contrast, a multi-floor dialogue consists of multiple sets of dialogue participants, each conversing within their own floor, but also at least one multi-communicating member who is a participant of multiple floors and coordinating each to achieve a shared dialogue goal. For example, in a restaurant, a server communicates with customers to take their orders in the dining room (one floor) and talks with other workers in the kitchen (another floor) who prepare the customer's food. All the participants work toward the joint goal of providing the customer with their desired meals, however in this case, only the server participates in both floors, conveying orders from customer to kitchen and perhaps information about item availability or speed from kitchen back to customers. Another example is in military units, where soldiers follow their commander's orders, which are decided at headquarters. Such situations are quite common in the real world, where we have different dialogue floors for decision-making and actions based on decisions.

Identifying aspects of multi-floor dialogue structure can be critical for building cooperative applications that have to participate in multi-floor dialogues, for example collaborative navigation robots [Lukin *et al.*, 2018; Bonial *et al.*, 2018]. However, most existing studies on dialogue structure parsing addressed only single-floor dialogues. There are standard annotation schemes for both dialogue acts [Bunt *et al.*, 2012] and discourse relations [Prasad and Bunt, 2015] in single-floor dialogues. Some proposed models have parsed the dialogue structure. However, these schemes do not fully address the issues of dialogue structure in multi-floor dialogues. A previous work proposed an annotation scheme of dialogue structure on multi-floor dialogues [Traum *et al.*, 2018]. This scheme is based on two important aspects of dialogue structure: transaction units and the relations between utterances. A transaction unit clusters utterances from multiple participants and floors that contribute to achieving the initiating participant's intention. Relations link utterances to antecedents within the unit. However, there is no previous work on automatic dialogue structure parsing for multi-floor dialogue.

In this paper, we propose a first neural dialogue structure parser for multi-floor dialogue structure. Our proposed parser has an attention mechanism to predict structure across different floors. In the following sections, we describe the dialogue structure parsing task on multi-floor dialogue, an annotation scheme, and our target domain (Section 2). We describe our proposed system based on the end-to-end approach, which automatically identifies the dialogue structure of multi-floor dialogues by recurrent neural networks. The definitions of transaction units, antecedents, and relation-types are closely related to each other. We applied the attention mechanism and multi-task learning to improve the overall performance of the dialogue structure parser considering their characteristics (Section 3). We experimentally evaluated the dialogue structure parsing performance of our model using automatic metrics that focus on micro- and meso-level structures [Traum and Nakatani, 1999] in dialogues (Section 4). Our proposed model using multi-task learning improved the overall performance compared to models trained on single-task settings (Section 5). Finally, we conclude by describing the performance of our proposed model and discussing possible future directions (Section 6).

Table 1: Dialogue example of multi-floor dialogue

	Left Floor		Right Floor		Annotations		
#	**Commander**	**DM → Commander**	**DM → RN**	**RN**	**TU**	**Ant**	**Rel**
1	move to where you see the first cone				1		
2		I ' m not sure which object you are referring to. Can you describe it in another way, using color or its location?			1	1	request-clarification
3	move to the cone on the right a red cone on the right				1	2	clarification-repair
4			move to face the cone on the right		1	3	translation-r
5		executing...			1	3	ack-doing
6	take another picture				2		
7				done	1	4	ack-done
8		done			1	7	translation-l
9			image		2	6	translation-r
10				image sent	2	9	ack-done
11		sent			2	10	translation-l

2 Dialogue Structure in Multi-floor Dialogue

For our initial investigations, we use a dataset of multi-floor dialogue structure, created as part of a long-term project to develop an autonomous robot [Marge *et al.*, 2016; Lukin *et al.*, 2018; Gervits *et al.*, 2019], which is commanded by remote human participants. The robot is in an unfamiliar physical environment, where it performs object searches through natural language interaction. The dataset consists of "Wizard of Oz" dialogues where two wizards control the robot and communicate with the human commander. The dialogue manager wizard (DM) communicates directly with the commander in natural language and handles clarifications or misconceptions that might not be applicable given the environment and robot capabilities. A robot navigator wizard (RN) controls the robot with a joystick controller, but communicates only with the DM. There are thus two separate floors - one between commander and "robot" (actually the DM), and one between the two wizards. These floors are called "left" and "right", for convenience. Table 1 shows an example of an actual dialogue excerpt, including two floors and four distinct message streams. The commander gives its intention to the DM on their dialogue floor (left floor). The DM talks with the commander (when necessary) to clarify the commander's intention. After completely understanding the commander's intention, the DM moves to another dialogue floor (right floor) to transfer the commander's intention to the RN, which operates the robot based on the given intention and reports the result to the DM. The DM returns to the first floor to feedback the result to the commander. Note that the DM can communicate with any participants by moving among several dialogue floors to transfer the information as a multi-communicator [Reinsch Jr *et al.*, 2008]; but the RN and the commander cannot directly communicate.

Previous work defined an annotation scheme for such multi-floor dialogues to specify their characteristics [Traum *et al.*, 2018]. To capture the information update process of the dialogue participants, this scheme focused on the intentional structure [Grosz and Sidner, 1986], which consists of units of multiple consecutive utterances, and the relations between pairs of utterances within the unit. They defined an annotation scheme for (1) transaction units, (2) antecedents, and (3) relation-types, and the dataset includes human-annotated data. In this study, we explore a model that automatically identifies these structures. Below we describe the annotation scheme in [Traum *et al.*, 2018].

2.1 Transaction Unit

A transaction unit (TU) is a basic unit of intentional structure in a multi-floor interaction. It consists of the initial utterance that expresses the intention of the speakers and every subsequent utterance across all the floors to achieve the original speaker's intention. Each utterance belongs to a transaction unit, which is defined by a set of utterances. The "TU" column of Table 1 shows an numerical identifier for the unit which is the same for all utterances that are part of the TU.

In some cases, multiple transactions are "active" at the same time, in that they have been initiated but not terminated. For example, Table 1 shows a case where two transaction units are included in the dialogue: TU1 is about moving somewhere, while TU2 is about taking a picture. TU2 is initiated in utterance #6, before TU1 is completed in utterance #8. Both transactions are thus running in parallel during this part of the dialogue.

2.2 Antecedent and Relation-Type

In [Traum *et al.*, 2018], relations are annotated between utterances in the same TU, using antecedents and relation-types. Any utterances after the first utterance in the transaction unit have antecedents, shown in the "Ant" column of Table 1, as the utterance ID of the antecedent utterance. Relation types are summarized in Table 2. These relations are categorized first as to whether they are from the same participant (expansions), from different participants in the same floor, or across floors. Each of these categories has a set of specific relations and in some cases sub-types. Relation types are indicated in the "Rel" column in Table 1.

The set of relations within a transaction define a tree structure, where the first utterance is the root node, which has no relation-type or antecedent annotations. In the example in Table 1, #1 and #6 are the root nodes of the two transaction units.

Table 2: Relation-types in a multi-floor dialogue

Type	Sub-types
Expansions	relate utterances that are produced by the same participant within the same floor.
	continue
	link-next
	correction
	summarization
Responses	relate utterances by different participants within the same floor.
	acknowledgment
	done
	doing
	wilco
	understand
	try
	unsure
	can't
	clarification
	req-clar
	clar-repair
	missing info
	nack
	repeat
	processing
	question-response
	answer
	non-answer
	other
	3rd turn feedback
	reciprocal response
Translations	relate utterances in different floors.
	transalation-l
	transalation-r
	comment
	quotation

3 Neural dialogue structure parser for Multi-Floor Dialogue

In this section, we introduce a neural dialogue structure parser for the annotation scheme proposed by [Traum *et al.*, 2018]. A dialogue structure parser based on end-to-end neural networks improved the parsing performance more than legacy models using hand-crafted features [Afantenos *et al.*, 2015; Shi and Huang, 2019]. Thus, we built an end-to-end neural dialogue structure parser model with available data and explored its limitations.

In our dialogue structure parsing task on multi-floor dialogues, three tasks are closely related: transaction units, antecedents, and relation-type identifications. We expect that multi-task learning will improve the overall parsing performance more than single models. The attention mechanism can explicitly represent their relations. Thus, our model is based on a recurrent neural network that has both soft and hard attention mechanism with multi-task learning.

Our proposed model (Fig. 1) mainly includes four networks:

- **Hierarchical encoder** has utterance and context encoders for encoding each dialogue context in different dialogue levels.
- **The antecedent predictor** estimates the antecedent that corresponds to each utterance.
- **The transaction-unit predictor** estimates the type of transaction boundaries of each utterance.
- **The relation-type predictor** estimates the relation-type of each utterance and its antecedent.

The transaction-unit and relation-type predictors share the prediction results of the antecedent predictor as attention weights, because their prediction results are related to the potential tree structures decided by the antecedent predictor model. Such a two-stage approach, which predicts the dependency structure of the utterances and its relation-types, resembles previous work [Shi and Huang, 2019]. However, that model targets single-floor dialogue structure parsing, and our model predicts the dialogue structure of multi-floor dialogues and clusters the utterances in different floors as one transaction unit.

3.1 Hierarchical Encoder

Our hierarchical encoder consists of utterance and context encoders. The utterance encoder receives a word at each time step using forward and backward GRUs [Cho *et al.*, 2014] to encode each utterance into a fixed-length vector:

$$\overrightarrow{h_{t,i}} = \overrightarrow{\text{GRU}}_{\text{utt}}(\overrightarrow{h_{t,i-1}}, \text{Embedding}(w_{t,i})), \quad (1)$$

$$\overleftarrow{h_{t,i}} = \overleftarrow{\text{GRU}}_{\text{utt}}(\overleftarrow{h_{t,i+1}}, \text{Embedding}(w_{t,i})), \quad (2)$$

$$h_{t,i} = [\overrightarrow{h_{t,i}}; \overleftarrow{h_{t,i}}], \quad (3)$$

$$h_t = \frac{1}{|U_t|} \sum_{i=1}^{|U_t|} h_{t,i}. \quad (4)$$

Here t is the utterance numbers in the dialogue context and i is the word order in the utterance. $h_{t,i}$ is the hidden vector calculated from each word $w_{t,i}$ and the hidden vector in previous

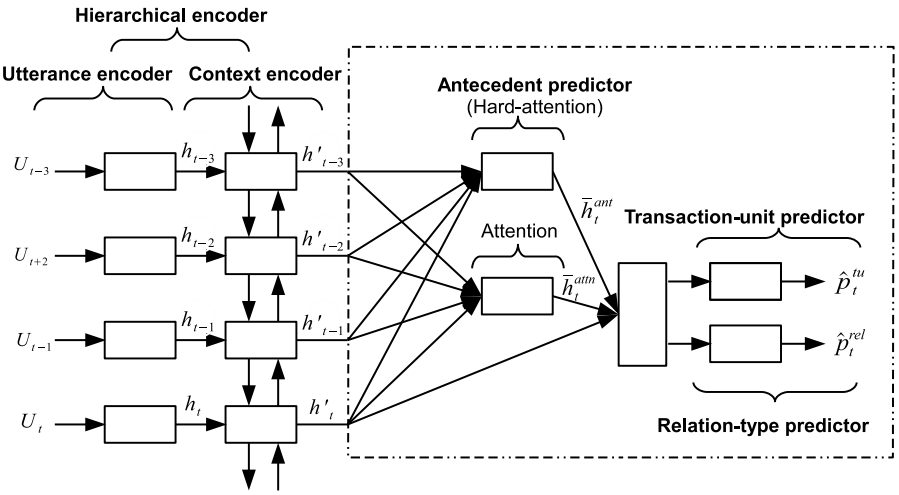

Figure 1: Overview of proposed neural dialogue structure parser

time-step $h_{t,i-1}$ in utterance $U_t = \{w_{t,1}, w_{t,2}, \cdots, w_{t,N}\}$. Each word $w_{t,i}$ is converted to a fixed-length vector using an embedding layer before calculating the hidden vector. In each utterance, we added a special symbol, which indicates the types of floors, to prefixes and suffixes of utterance and trained the embedding rule as done with words.

In the context encoder, utterance vectors are input to encode the dialogue history to get context-level vector representation h'_t for each utterance in the dialogue contexts:

$$\overrightarrow{h'_t} = \overrightarrow{\text{GRU}}_{\text{hist}}(\overrightarrow{h'_{t-1}}, h_t)), \tag{5}$$

$$\overleftarrow{h'_t} = \overleftarrow{\text{GRU}}_{\text{hist}}(\overleftarrow{h'_{t+1}}, h_t), \tag{6}$$

$$h'_t = [\overrightarrow{h'_t}; \overleftarrow{h'_t}]. \tag{7}$$

We introduce a soft-attention mechanism [Luong *et al.*, 2015] for dialogue contexts to compute contextual representation $\bar{h}^{\text{attn}}_t$ for each utterance U_t:

$$\text{attention}(h'_{t-j}, h'_t) = h'^T_{t-j} W_{\text{ant}} h'_t, \tag{8}$$

$$\alpha_j = \frac{\exp(\text{attention}(h'_{t-j}, h'_t))}{\sum_{j=1}^{k} \exp(\text{attention}(h'_{t-j}, h'_t)}, \tag{9}$$

$$\bar{h}^{\text{attn}}_t = \sum_{j=1}^{k} \alpha_j \cdot h'_{t-j}. \tag{10}$$

Here k is the number of previous utterances considered in the calculation of attention, W_{attn} is a trainable weight-matrix, and $\alpha_j \in [0,1]^k$.

In addition, we introduce a hard-attention mechanism for explicitly considering the antecedent, which corresponds to each turn t:

$$\bar{h}^{\text{ant}}_t = \sum_{j=1}^{k} \beta_j \cdot h'_{t-j} \tag{11}$$

Here β_j takes 1 if utterance U_{t-j} is the antecedent of utterance U_t and 0 in other cases ($\beta_j \in \{0,1\}^k$).

Attention vectors $\bar{h}^{\text{attn}}_t$ and $\bar{h}^{\text{ant}}_t$, which are calculated on the basis of the hard and soft-attention mechanisms, are combined:

$$\hat{h}^{\text{fc}}_t = \tanh(\text{Linear}_{\text{attn}}([\bar{h}^{\text{attn}}_t; \bar{h}^{\text{ant}}_t; h'_t]). \tag{12}$$

Here $\text{Linear}_{\text{attn}}$ is a linear transformation layer, which includes a bias term. $\hat{h}^{\text{fc}}_t$ is a shared vector for predicting the transaction units and relation-types. Note that gold antecedent β is used in training; however, in the inference, the model uses predicted distribution of $\bar{h}^{\text{ant}}_t$ by the antecedent predictor.

3.2 Antecedent Predictor

As shown in Table 1, each utterance has an annotation of the corresponding antecedent as its utterance ID (#). To predict the antecedents for each utterance U_t, we calculated the scores between each utterance and the contextual utterances:

$$\text{antecedent}(h'_{t-j}, h'_t) = h'^T_{t-j} W_{\text{ant}} h'_t, \tag{13}$$

$$\hat{\beta}_j = \frac{\exp(\text{antecedent}(h'_{t-j}, h'_t))}{\sum_{j=1}^{k} \exp(\text{antecedent}(h'_{t-j}, h'_t)}. \tag{14}$$

Here, k is the number of preceding utterances that can be the antecedent, W_{ant} is a trainable weight-matrix, and $\hat{\beta}_j \in [0,1]^k$. By calculating the position of antecedent from the weights of attention, we can carry this knowledge forward to other predictions in the rater step: transaction-unit prediction and relation-types prediction.

We set the cross-entropy loss between predicted distribution $\hat{\beta}$ and actual antecedent label β as a loss function that enforces that the contextual utterance has the highest score when it is the antecedent of U_t:

$$L_{t,\text{ant}} = -\sum_{j=1}^{k} \beta_j \log(\hat{\beta}_j). \tag{15}$$

Note that we also calculate the attention weight corresponding to the case where the utterance does not have any antecedent using the trainable vector and the hidden vector h'_t.

3.3 Transaction-Unit Predictor

We formulate the problem of transaction unit prediction as a sentence classification problem that determines the boundaries of the transaction units in dialogues. The transaction-unit predictor classifies each utterance into the following three classes:

- **Start**: the utterance is the beginning of a transaction unit.
- **Continue**: the utterance belongs to the same transaction unit as the previous utterance.
- **Other**: the utterance cannot be categorized into either of the above classes.

Other indicates that the utterance belongs to an already open transaction that is different from the one the previous utterance belongs to, such as utterance #7 and #9 in Table 1. We predict transaction boundaries using $\hat{h}_t^{\text{fc}}$, derived from the calculation results of soft and hard-attentions to the context:

$$\hat{p}_t^{\text{tu}} = \text{softmax}(\text{Linear}_{\text{tu\_pred}}(\hat{h}_t^{\text{fc}})). \quad (16)$$

Here $\text{Linear}_{\text{tu\_pred}}$ is a linear transformation layer that includes a bias term, and $\hat{p}_t^{\text{tu}}$ is the predicted distribution of the transaction boundaries.

We used the cross-entropy loss as the loss function:

$$L_{t,\text{tu}} = -\sum_{j=1}^{|p_t^{\text{tu}}|} p_j^{\text{tu}} \log(\hat{p}_{t,j}^{tu}). \quad (17)$$

Here p_t^{tu} is a three-dimensional vector corresponding to the type of target transaction boundaries.

3.4 Relation-Type Predictor

We used $\hat{h}_t^{\text{fc}}$ as well as the transaction-unit predictor to predict the relation-type of each utterance with its antecedent:

$$\hat{p}_t^{\text{rel}} = \text{softmax}(\text{Linear}_{\text{rel\_pred}}(\hat{h}_t^{\text{fc}})). \quad (18)$$

Here $\text{Linear}_{\text{rel\_pred}}$ is a linear transformation layer that includes the bias term and $\hat{p}_t^{\text{rel}}$ is the predicted distribution of the relation-types.

We used the cross-entropy loss for the training:

$$L_{t,\text{rel}} = -\sum_{j=1}^{|p_j^{\text{rel}}|} p_j^{\text{rel}} \log(\hat{p}_{t,j}^{rel}). \quad (19)$$

Here p_t^{rel} is a vector, whose dimensions correspond to a relation label defined in Table 2.

3.5 Objective Function

We have to optimize the above three models not only to a single model but also to the other two models because these tasks are closely related. In this study, we introduce a multi-task loss, which combines each prediction loss of the antecedent, the transaction-unit, and the relation-type predictor. In multi-task learning, we interpolate the loss functions of three tasks:

$$L = \frac{1}{N}\sum_{t=1}^{N}(\gamma_{\text{ant}} L_{t,\text{ant}} + \gamma_{\text{tu}} L_{t,\text{tu}} + \gamma_{\text{rel}} L_{t,\text{rel}}). \quad (20)$$

Here N is the dialogue length. γ_{ant}, γ_{tu}, and γ_{rel} are the weights for adjusting the importance of each predictor in the loss calculation.

4 Experimental Settings

In our experiment, we evaluated the dialogue structure parsing performance of our proposed model. In this section, we describe the dataset for the training and evaluation, the setting of the model training, and the evaluation metrics.

4.1 Dataset

We used a dataset [Traum *et al.*, 2018] that contains Exp. 1 and Exp. 2 data[1]. The dialogues were annotated based on a previously described scheme [Traum *et al.*, 2018], which was specifically designed to handle multiple conversational floors. As shown in Table 3, these dialogue data consist of 48 dialogues (1829 transactions) executed by several different commanders.

Table 3: Numbers of dialogues, utterances, and transactions

	Dialogues	Utterances	Transactions
Exp. 1	24	4527	780
Exp. 2	24	6994	1049

To evaluate the parsing performance of the proposed model, we randomly divided all of the dialogues in Exp. 1 and Exp. 2 into six subsets and applied double cross-validation [Mosier, 1951]. We used a single subset for validation and a test-set for each, and the remaining subset was used as training data. We evaluated every possible combination of training, validation, and test-set and the final performance by a majority vote on the prediction results of the models, which share the same test-set.

4.2 Model Settings

We evaluated the dialogue structure parsing performance of the proposed model in multi-floor dialogues by comparing the dialogue structure parsing performances of the proposed model with the multi-task loss (**Multi**) and the models individually trained for each task (**Single**). We also compared the cases based on both the **Offline** and **Online** models. The proposed model described in Section 3 uses bi-directional GRUs in the context encoder to make predictions for each utterance U_t; this means the model cannot start parsing during the dialogue. We call this setting **Offline**. In contrast, we also considered a model that only uses previous contexts without subsequent contexts in the prediction for each utterance U_t. We call this setting **Online**. The online model is important for real-time dialogue robot processing, which can only use the observed information based on the interaction sequence. We built the online model only using forward-GRUs instead of bidirectional-GRUs in the context encoder.

We used the same hyper-parameter settings in each model. The vocabulary size was 500, the word embedding size was 100, and the hidden size was 300. We used byte pair encoding (BPE) for tokenization [Sennrich *et al.*, 2016]. In training, we used a mini-batch size of 64 and an Adam optimizer [Kingma and Ba, 2014] with a learning rate of 1e-4. We set γ_{ant}, γ_{tu}, and γ_{rel} to 1. In the relation-type prediction, we integrated the 'acknowledgement," "clarification,"

[1]The annotation for Exp. 3 is still in progress.

Table 4: Prediction performances of transaction units, antecedents, and relation-types

Models	TU				Ant				Rel			
	$\overline{\text{Prec.}}$	$\overline{\text{Rec.}}$	$\overline{\text{F1}}$	TuAcc	$\overline{\text{Prec.}}$	$\overline{\text{Rec.}}$	$\overline{\text{F1}}$	TreeAcc	$\overline{\text{Prec.}}$	$\overline{\text{Rec.}}$	$\overline{\text{F1}}$	TreeAcc w/ rel
Majority	55.71	74.64	63.80	-	23.59	48.57	31.76	-	8.54	29.22	13.21	-
Single-Online	95.43	95.46	95.44	81.19	93.92	90.89	92.34	68.12	92.38	92.84	92.53	63.80
Multi-Online	95.99	95.99	95.99	84.25	93.93	90.84	92.33	70.09	93.69	94.11	93.80	66.81
	(96.26)	(96.27)	(96.26)	(85.34)	-	-	-	-	(94.75)	(94.94)	(94.77)	(67.74)
Single-Offline	95.31	95.35	95.33	81.46	94.26	90.70	92.40	68.83	94.86	93.22	92.91	64.62
Multi-Offline	**96.05**	**96.07**	**96.06**	**84.52**	**94.58**	**91.95**	**93.21**	**71.35**	**93.75**	**94.26**	**93.90**	**69.05**
	(96.30)	(96.31)	(96.30)	(85.51)	-	-	-	-	(94.68)	(94.95)	(94.69)	(69.92)

and "question-response" sub-types into these classes because some sub-types rarely appeared in the dataset. In addition, we defined a label where utterance has no antecedent, as well as relation-types (#1 and #6 in Table 1).

4.3 Evaluation Metrics

We defined the micro and meso-level evaluation metrics for our dialogue structure parsing task. For the micro-level evaluation, we defined the label prediction performances of the antecedents, the transaction units, and the relation-types by precision (Prec.), recall (Rec.), and F1. Note that we took the relative position of each utterance from its antecedent as a label to compute the metrics when evaluating the antecedent prediction performance. In other words, we compared the difference between the position of predicted antecedents and actual antecedents. We also introduced metrics for the meso-level structure [Traum and Nakatani, 1999] in dialogues to evaluate the consistency of the parsing results. We used the following three metrics:

- **TuAcc** is the ratio of the transaction units that perfectly predicted the transaction boundaries for each utterance within the transaction unit.
- **TreeAcc** is the ratio of the transaction units that perfectly predicted the antecedents for each utterance within the transaction unit.
- **TreeAcc w/ rel** is the ratio of the transaction units that perfectly predicted the antecedents and the relation-types for each utterance within the transaction unit.

Note that the meso-level metrics are stricter than the micro-level metrics, which judge the prediction result of each utterance.

5 Experimental Results

Table 4 shows the performances of each dialogue structure parser. Here **Single** denotes a case where the transaction unit, the antecedent, and the relation-type predictors were individually trained. **Multi** denotes a case where these models were trained with multi-task learning loss. In addition, **Offline** indicates that the model used the dialogue entirely for parsing, and **Online** indicates that the model used only the preceding context of each utterance. **Majority** denotes a case where the always predicts the most frequent label. $\overline{\text{Prec.}}$, $\overline{\text{Rec.}}$, and $\overline{\text{F1}}$ are the weighted averages[2] of the precision, recall, and F1 scores of the predicted labels. The brackets are the prediction results where the oracle antecedent was fed into the model.

The result shows that the dialogue structure parsing performance (transaction unit prediction, antecedent prediction, and relation-types prediction) of the **Offline** models have slightly improved from the **Online** models. This result indicates that subsequent contexts are useful for predicting labels for each utterance, but we have enough prediction accuracy without using the subsequent context. When the **Multi** models use oracle antecedents to predict the transaction units and the relation-type, we can further improve the performance of dialogue structure parsing.

Table 5: Transaction unit prediction performance of Multi-Offline model

Tu-Label	Precision	Recall	F1	Count
Start	0.9466	0.9306	0.9385	1829
Continue	0.9750	0.9800	0.9775	8599
Other	0.8694	0.8591	0.8642	1093
Weighted-Avg	0.9605	0.9607	0.9606	11521

Table 5 shows the results of the transaction unit prediction in **Multi-Offline** model, which was the best model in our experiment. We confirmed that the model achieved over 85% F1 for all labels (types of transaction boundary). However, there is still a problem in the predicting "Other" (when the utterance belongs to a different transaction than the previous utterance, but it is not the start of a new transaction). Our model decided labels with the highest prediction probability for each utterance; however, we did not take into account the consistency of the prediction results in the sequence of dialogue. To solve this problem, we can introduce a model that takes into account information about the entire prediction results, such as Conditional Random Field (CRF) [Lafferty *et al.*, 2001] for further improvements.

Table 6 shows the results of the antecedent prediction in the **Multi-Offline** model. Here each label indicates the relative position from each utterance to its antecedent. Note that this table only shows the prediction results by considering a maximum of 10 previous utterances. We excluded a few cases

[2]We calculated the weighted averages based on the label frequencies.

Table 6: Antecedent prediction performance of Multi-Offline model

Position	Precision	Recall	F1	Count
-10	58.33	66.67	62.22	21
-9	73.08	57.58	64.41	33
-8	82.50	63.46	71.74	52
-7	80.77	74.12	77.30	85
-6	90.48	85.39	87.86	178
-5	81.31	64.44	71.90	135
-4	91.87	85.32	88.48	477
-3	94.98	87.00	90.82	1023
-2	94.84	93.06	93.94	2509
-1	95.99	95.47	95.73	4262
Weighted-Avg	94.58	91.15	93.21	8775

when the antecedent of the utterance is not included in the ten previous utterances, and the utterance has no antecedent. Our model can predict antecedents with high performance when the relative position was not distant. On the other hand, the prediction performance was below 80% when the relative positions were distant (greater than five in absolute). This result suggests the difficulty of addressing long-term dependency in dialogues. In addition, our model ignores the consistency of the tree structure associated with the predicted antecedents. The search for dialogue structures using dynamic programming probably has the potential to improve the performance of our model.

Table 7: Relation-type prediction performance of Multi-Offline model

Relation-Type	Precision	Recall	F1	Count
Expansions				
-continue	0.9090	0.8890	0.8989	955
-link-next	0.9937	0.9969	0.9953	318
-correction	0.4000	0.1111	0.1739	36
-summarization	0.00	0.00	0.00	13
Responses				
-acknowledgement	0.9729	0.9706	0.9717	3366
-clarification	0.7951	0.8500	0.8216	420
-processing	1.0000	0.9957	0.9978	233
-question-answer	0.5705	0.5174	0.5427	172
-other	0.3333	0.0606	0.1026	33
-3rd-turn-feedback	0.5000	0.0400	0.0741	25
-reciprocal-response	0.00	0.00	0.00	5
Translations				
-l	0.9593	0.9814	0.9703	1563
-r	0.9830	0.9840	0.9835	1942
-comment	0.4000	0.3810	0.3902	21
No-antecedent	0.9185	0.9463	0.9322	2419
Weighted-Avg	0.9375	0.9426	0.9390	11521

Table 7 shows the results of the relation-type predictions in the **Multi-Offline** model. Our model showed higher F1 scores in frequent relation-types, including when the utterance has no antecedent. There is still a challenge in predicting low-frequent relation-types due to the lack of training data. Ongoing annotation work [Traum *et al.*, 2018] with additional data may remedy this problem. We also need to look at ways to deal with these unbalanced labels.

Finally, in Table 8, we show an example of dialogue structure parsing on a fragment of multi-floor dialogue. Note that we displayed the correct labels in brackets when the label was incorrectly predicted, and "#" corresponds to cases where the utterance does not have the antecedent. The first example shows that the model accurately predicts all the transaction boundaries, antecedents, and relation-types, even if transactions were interleaved. However, the second example is including error predictions of transaction boundaries. In this example, there are only two TUs, but the model has determined that the utterance has three TUs. Note that, even if we assume the prediction of TU at #8 is correct, the prediction at #11 is still not correct. When such confusion occurs, the error extends beyond one utterance to multiple utterances. In many cases, delays in communication and differences in the quality of annotations between Exp.1 and 2 often confuse predictions.

6 Conclusion

We built a neural dialogue structure parser with an attention mechanism that applies multi-task learning to automatically identify the dialogue structure of multi-floor dialogues. The experimental results showed that our proposed model improved the identification performance on all tasks compared to the model trained on single task settings. However, problems remain with the performance of the dialogue structure identification due to the lack of training data, especially for rare labels. To prevent this problem, we will consider pre-training and the transfer learning of models using existing dialogue corpora and discourse-relation datasets. We also explore the possibility of introducing powerful models of similar tasks related for predicting tree-structure in a document, such as a dependency parsing [Nivre, 2010] and discourse parsing based on rhetorical structure theory [Webber *et al.*, 2012; Mann and Thompson, 1988].

This study has developed the first baseline model for automatic identification of dialogue structure on multi-floor dialogues. It has a potential for applying to the automatic annotation of dialogue structure on multi-floor dialogues, and encourage the development of a dialogue manager and robot navigator on multi-floor settings.

Acknowledgments

This research and development work was supported by Grant-in-Aid for JSPS Fellows No. 20J14823. We would like to thank United States Army Research Laboratory for sharing the dataset for this research. Author Traum was supported in part by the Army Research Office under Cooperative Agreement Number W911NF-20-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

Table 8: Examples of the dialogue structure parsing on multi-floor dialogue

	Left Floor		Right Floor		Prediction		
#	**Commander**	**DM → Commander**	**DM → RN**	**RN**	**TU**	**Ant**	**Rel**
1	turn right twenty degrees				Start	#	#
2			turn right 20		Continue	1	translation-r
3		executing ...			Continue	1	response-ack.
4			image		Continue	1	translation-r
5				done image sent	Continue	4	response-ack.
6	go forward fifteen feet				Start	#	#
7		sent			Other	5	translation-l
8	and go through door on right				Other	6	expansion-cont.
9			move forward about 15 feet , going through door on right , image		Continue	8	translation-r
10		executing ...			Continue	8	response-ack.
1	take a picture				Start	#	#
2			image		Continue	1	translation-r
3				image sent	Continue	2	response-ack.
4		sent			Continue	3	translation-l
5	turn left ninety degrees				Start	#	#
6			turn left 90		Continue	5	translation-r
7		executing ...			Continue	5	response-ack.
8	take a picture after each command				Start (**Continue**)	# (**5**)	# (**expansion-cont.**)
9				done	Other (**Continue**)	6	response-ack.
10			take pic after each command		Other (**Continue**)	8	translation-r
11			image		Other (**Continue**)	8	translation-r
12				image sent	Continue	11	response-ack.
13		sent			Continue	12	translation-l

References

[Afantenos *et al.*, 2015] Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[Bonial *et al.*, 2018] Claire Bonial, Stephanie M Lukin, Ashley Foots, Cassidy Henry, Matthew Marge, Kimberly A Pollard, Ron Artstein, David Traum, and Clare R Voss. Human-robot dialogue and collaboration in search and navigation. In *Proceedings of the Annotation, Recognition and Evaluation of Actions (AREA) Workshop of the 2018 Language Resources and Evaluation Conference (LREC)*, 2018.

[Bunt *et al.*, 2012] Harry Bunt, Jan Alexandersson, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Volha Petukhova, Andrei Popescu-Belis, and David R Traum. ISO 24617-2: A semantically-based standard for dialogue annotation. In *Proc. of Conference on International Language Resources and Evaluation*, pages 430–437, 2012.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *Proc. of EMNLP*, pages 1724–1734, 2014.

[Gervits *et al.*, 2019] Felix Gervits, Anton Leuski, Claire Bonial, Carla Gordon, and David Traum. A Classification-Based Approach to Automating Human-Robot Dialogue. In *Proceedings of International Workshop on Spoken Dialog System Technology (IWSDS)*, page 12, Siracusa, Italy, April 2019.

[Grosz and Sidner, 1986] Barbara J Grosz and Candace L Sidner. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204, 1986.

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

[Lafferty *et al.*, 2001] John Lafferty, Andrew McCallum, and Fernando CN Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[Lukin *et al.*, 2018] Stephanie Lukin, Felix Gervits, Cory Hayes, Pooja Moolchandani, Anton Leuski, John G Rogers III, Carlos Sanchez Amaro, Matthew Marge, Clare Voss, and David Traum. Scoutbot: A dialogue system for collaborative navigation. In *Proceedings of ACL 2018, System Demonstrations*, 2018.

[Luong *et al.*, 2015] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[Mann and Thompson, 1988] William C Mann and Sandra A Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[Marge *et al.*, 2016] Matthew Marge, Claire Bonial, Kimberly A Pollard, Ron Artstein, Brendan Byrne, Susan G Hill, Clare Voss, and David Traum. Assessing agreement in human-robot dialogue strategies: A tale of two wizards. In *International Conference on Intelligent Virtual Agents*, pages 484–488. Springer, 2016.

[Mosier, 1951] Charles I Mosier. I. problems and designs of cross-validation 1. *Educational and Psychological Measurement*, 11(1):5–11, 1951.

[Nivre, 2010] Joakim Nivre. Dependency parsing. *Language and Linguistics Compass*, 4(3):138–152, 2010.

[Prasad and Bunt, 2015] Rashmi Prasad and Harry Bunt. Semantic relations in discourse: The current state of iso 24617-8. In *Proceedings of the 11th Joint ACL-ISO Workshop on Interoperable Semantic Annotation (ISA-11)*, 2015.

[Reinsch Jr *et al.*, 2008] N Lamar Reinsch Jr, Jeanine Warisse Turner, and Catherine H Tinsley. Multicommunicating: A practice whose time has come? *Academy of Management Review*, 33(2):391–403, 2008.

[Sennrich *et al.*, 2016] Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proc. ACL*, volume 1, pages 86–96, 2016.

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

[Traum and Nakatani, 1999] David Traum and Christine H Nakatani. A two-level approach to coding dialogue for discourse structure: activities of the 1998 dri working group on higher-level structures. In *Towards Standards and Tools for Discourse Tagging*, 1999.

[Traum *et al.*, 2018] David Traum, Cassidy Henry, Stephanie Lukin, Ron Artstein, Felix Gervits, Kimberly Pollard, Claire Bonial, Su Lei, Clare Voss, Matthew Marge, et al. Dialogue structure annotation for multi-floor interaction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[Webber *et al.*, 2012] Bonnie Webber, Markus Egg, and Valia Kordoni. Discourse structure and language technology. *Natural Language Engineering*, 18(4):437, 2012.