GrounDialog: A Dataset for Repair and Grounding in Task-oriented Spoken Dialogues for Language Learning

Xuanming Zhang[1*], **Rahul Divekar**[2], **Rutuja Ubale**[2], and **Zhou Yu**[1]

[1]Computer Science Department, Columbia University
[2]AI Research Labs, Educational Testing Service
{xz2995,zy2461}@columbia.edu
{rdivekar,rubale}@ets.org

Abstract

Improving conversational proficiency is a key target for students learning a new language. While acquiring conversational proficiency, students must learn the linguistic mechanisms of Repair and Grounding (R&G) to negotiate meaning and find common ground with their interlocutor so conversational breakdowns can be resolved. Task-oriented Spoken Dialogue Systems (SDS) have long been sought as a tool to hone conversational proficiency. However, the R&G patterns for language learners interacting with a task-oriented spoken dialogue system are not reflected explicitly in any existing datasets. Therefore, to move the needle in Spoken Dialogue Systems for language learning we present GrounDialog: an annotated dataset of spoken conversations where we elicit a rich set of R&G patterns.

1 Introduction and Motivation

Many conversations are impromptu back-and-forth interactions that often have no prior preparation or review. As a result, conversational breakdowns (Benner et al., 2021; Li et al., 2020) may occur due to minor misinterpretation, mishearing, misspeaking, or a general lack of common ground (Traum, 1994). Interlocutors use Repair mechanisms (Albert and de Ruiter, 2018) to detect and resolve communicative problems during conversations; and Grounding mechanisms to establish common ground. For example, we often ask our interlocutors to repeat what they said, explain themselves, request clarifications, etc. Such processes arise proactively or when the initial communication attempt has failed, during which modification and revision to the previous utterances are needed to proceed the conversations naturally.

According to Long (1983), R&G is meaningful in the following perspectives: 1) repair the discourse when breakdown occurs and 2) avoid conversational breakdowns. Table 1 shows an example dialogue between low-proficiency (LPS) and high-proficiency (HPS) English speakers, where LPS paraphrases themselves to repair the discourse when trouble occurs. Besides, speakers usually try their best to avoid breakdowns in conversations. Based on Long (1983), there are plenty of strategies they can adopt to prevent the breakdowns during communications: 1) relinquish topic control; 2) simplify topic by asking "yes-no" questions; 3) confirm comprehensions of speakers before proceeding, etc.

Speaker	Transcriptions
LPS	What um what types presentation is expected?
HPS	I did not understand your last question. Can you be clear?
LPS	**I mean what types of presentation would you uh would you expected during the interview?**
HPS	You can put up a formal presentation based on your educational background.

Table 1: Example dialogue from GrounDialog. LPS stands for Low-Proficiency Speakers, whereas HPS represents High-Proficiency Speakers.

From the perspective of a language learner, dialogues serve as important media in language acquisition and learning (Eszenyi and van der Wijst, 2006). When language learners chat with high-proficiency speakers, language learners make considerable efforts to ground what they have to say (Eszenyi and van der Wijst, 2006). More specifically, the low-proficiency speakers (LPS) attempt to negotiate the meanings of conversations with high-proficiency speakers (HPS). According to Foster and Ohta (2005) and Cook (2015), interactional processes including negotiation for meaning and various kinds of repair and grounding are among the many ways learners gain access to the second language acquisition. Besides, LPS can also en-

[*]This work was done while first author was an intern at ETS.

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 300–314

hance their language skills, general communication skills and cultural knowledge during the conversations with HPS (Eszenyi and van der Wijst, 2006).

While R&G is common in nearly all conversations, it is particularly important for language learners as learners are still building up the full understanding of the language. They may also bring R&G influences of their primary language into the language they are learning. It is also possible that low-proficiency speakers (or language learners) employ additional or different R&G mechanisms than high-proficiency speakers of a language. Therefore, there is a lot to know about R&G mechanisms from low-proficiency speakers.

In this paper, we present a dataset that can help linguists and other researchers with several novel linguistic tasks such as identifying R&G patterns. Further, while repair and grounding is an important linguistic mechanism, it is rarely reflected explicitly in the design of spoken dialogue systems that aim to help people learn a new language. Our dataset can fill this gap by allowing researchers to model dialogue state tracking with R&G, generating responses with R&G turns, etc.

We collected this dataset by connecting a high-proficiency speaker and a low-proficiency speaker on a crowd sourcing platform. The high-proficiency speaker played the role of a human resources (HR) assistant in a wizard-of-oz style and was tasked to convey information about an interview. The low-proficiency speaker played the role of an interviewee and was tasked with finding specific information about the same interview through their conversation with the high-proficiency speaker. While R&G may occur as a course of natural conversation, we further induced it by giving the interlocutors some conflicting and incomplete pieces of information. We collected the voice of the low-proficiency speaker and the text responses of the wizard.

To the best of our knowledge, GrounDialog dataset is the first task-oriented dialogue dataset specifically tailored for repair and grounding in spoken conversations between high-proficiency and low-proficiency speakers. Each dialogue in the dataset is transcribed by human experts and contains vocal markers and disfluencies, such as "uh" and "um". It is annotated with R&G types, intents, and slots that are relevant to dialogue state mapping tasks. Hence, GrounDialog can be used to develop a task-oriented conversational agent, equipped with the R&G ability to detect communicative trouble, and adopt certain strategies to repair the discourse when trouble occurs.

The rest of the paper presents related work, details of the data collection process, the data annotation scheme, analyses of the data, and initial model benchmarks.

2 Related Work

As indicated in Dorathy and Mahalakshmi (2011), task-based language teaching (TBLT) puts emphasis on the utilization of tasks as the critical element in the language classroom given that tasks can offer better contexts for active language acquisition and second language promotion. From the perspective of dialogue systems, it is the task-oriented dialogue (ToD) that can help language learners achieve their proficiency goals through task completion. Previous dialogue systems have shown great promise in increasing second language acquisition proficiency. Bibauw et al. (2019) provide an overview of all spoken dialogue systems for language learning. Timpe-Laughlin et al. (2022) have compared learning language via role-play with a spoken dialogue system versus human, and found that spoken dialogue systems are a feasible alternative to human interaction in the role-playing context. Divekar et al. (2021) have found that interaction with spoken dialogue systems in immersive contexts improved students proficiency and decreased their anxiousness while using a foreign language thereby indicating there may be increased willingness to communicate with automated humanoid interlocutors. All this points to evidence that spoken dialogue systems are an effective tool for language acquisition.

Many spoken dialogue systems for the use of language learning have been built using off-the-shelf intent and slot detectors, and dialogue state managers (Bibauw et al., 2019). Divekar et al. (2018) have found some repair and grounding mechanisms in their dialogue system for language learning such as systems being able to respond to learners' questions like "what do you mean" or "what can I say next" in a rule-based system. However, quick scaling up for such systems can only come with datasets.

Several datasets exist to help build task-oriented dialogues such as Schema-Guided-Dialogue (SGD) (Rastogi et al., 2020), MultiWoZ (Budzianowski et al., 2018), Dialogue State Tracking Challenges (DSTC) 1-3 (Williams et al., 2013; Henderson et al.,

2014a,b) and DSTC 4-5 (Kim et al., 2017). Besides, there are other frequently used speech-based ToD data, including Fluent Speech Commands (FSC) [1], Audio-Snips (Coucke et al., 2018), Carnegie Mellon Communicator Corpus (CMCC)(Bennett and Rudnicky, 2002) and Let's Go Dataset [2].

However, existing task-oriented dialogue datasets do not reflect the language learning perspective as there are no constraints in their collection process that one interlocutor must be a low-proficiency speaker. Moreover, most datasets are also a result of a text-based interaction (Wang et al., 2019; Chen et al., 2021; Liang et al., 2021). This also means that the existing datasets will not contain R&G patterns specific for language learners interacting with a task-oriented *spoken* dialogue systems.

Therefore, we present a new dataset, namely GrounDialog, which will be the first dedicated ToD dataset specifically tailored for R&G in HPS-LPS conversations. Besides, the dataset can address the need for R&G in spoken form in specific scenarios that do not exist in the text-based exchange.

3 Data Collection Set-up

Our goal was to collect conversations between high-proficiency (HPS) and low-proficiency speakers (LPS). To accomplish this, we use Amazon Mechanical Turk (AMTurk) to recruit and connect pairs of HPS and LPS for our study. To identify whether a participant is HPS or LPS, we provided the participants descriptions of CEFR levels (Council of Europe, 2001) and asked them to self-identify their proficiency level [3]. For the purposes of this study, turkers who identify themselves as *Beginner, Elementary, Intermediate,* and *Upper Intermediate* i.e., A1-B2 levels were regarded as LPS; whereas those selecting *Advanced* and *Proficient* i.e., C1-C2 are considered as HPS. An assumption of our study is that we draw the line between HPS and LPS arbitrarily at B2 and trust the turker's self-reported proficiency to be accurate. With this setup, we can end up with nearly equal size of HPS and LPS, which can ease the turker-pairing process for our data collection. A detailed explanation of the data collection process and conversational task for both HPS and LPS is shown below. Subsequently, we will present the general statistics of the collected dialogues and users. The study was approved by the IRB of the institute conducting this research. All participants were adults and provided consent before starting data collection. All collected data released with the paper is anonymized to our best abilities.

3.1 Conversational Task

In order to collect the conversational data that fits our purpose of having a conversation between an automated interlocutor and human, we follow the Wizard-of-Oz set-up (Kelley, 1984). The set-up has also been validated by many previous studies (Wen et al., 2016; Asri et al., 2017; Budzianowski et al., 2018). In general, two turkers (i.e. one HPS and the other LPS) were paired to communicate with each other. We contextualize their task into a pre-interview setting, where an HR hiring manager talks to an interviewee. Specifically, we set LPS to be the interviewee and HPS to be the HR hiring manager. We assign different goals for each role: the interviewee needs to find out the answers to a set of interview-related questions (e.g. interview time, duration, location, etc.), whereas the HR manager is given the information LPS will need and asked to be in charge of scheduling an appointment with the connected interviewee. To induce more repair and grounding turns in the conversation, we provided overlapping but inconsistent information to the interlocutors. For example, the interviewee is instructed that the interview is going to be 30 minutes, whereas the HR manager has 45 minutes in their task specification. We assumed that the difference in information will lead to the interlocutors being confused, asking clarification from each other, and resolving the situation by picking a time (Foster and Ohta, 2005).

3.2 Dialogue Interface

To establish a stable live connection between two turkers, we adapted VisDial AMT Chat (Das et al., 2017) to connect two humans, enable voice input/output, and connect to an off-the-shelf text-to-speech service.

To simulate a Wizard-of-Oz like setting, we enable the LPS to directly record their speech, whereas the HPS input texts into a chat box and their responses are converted into speech using an off-the-shelf Text-To-Speech. The synthesized speech is played on the LPS side. In this way, the LPS could get a feeling of being connected to a

[1] https://fluent.ai/fluent-speech-commands-a-dataset-for-spoken-language-understanding-research/

[2] https://dialrc.github.io/LetsGoDataset/

[3] The complete pre-chat survey form is shown in appendix A

"chatbot", even though the responses are actually written by a human. The instructions for the LPS said that they will be connected to a human or a chatbot. In this way, we left it ambiguous for the LPS to decide for themselves whether they are talking to a chatbot or not. The HPS were told that they would appear as a bot so as to elicit bot-like communication from them. The example dialogue interfaces together with the instructions for both HPS and LPS are shown in Figure 8.

3.3 Data Statistics

In total, we collected 42 dialogues, including 1,569 turns, from 55 unique turkers, where there are 29 high-proficiency speakers (HPS) and 26 low-proficiency speakers (LPS). Dialogues collected in our dataset are fairly long, with an average number of 37.4 turns per dialogue. Figure 2 presents a distribution over the sentence lengths for both HPS and LPS. The average sentence lengths are 10.02 and 8.55 for HPS and LPS respectively. We collected a total of 793 spoken utterances from LPS, and 777 textual responses from HPS.

3.4 User Statistics

After completing the conversational task, we asked each turker to input their demographic information through a post-chat survey form [4].

Specifically, for the turkers who did fill in our survey after the chat, there are 35 males and 16 females, with the age spanning from 22 to 63. The majority of the turkers are from India (45%) and the Unitied States (37%). Also, the self-identified English proficiency levels based on CEFR (Council of Europe, 2001) for the collected users are shown in Figure 1. As mentioned before, we take C1-C2 as high-proficiency speaker, and A1-B2 as low-proficiency speaker.

3.5 Speech data and transcriptions

There are 793 audio recordings collected from the accepted LPS[5], of which 586 audio files are transcribed by SpeechPad [6], a reliable third-party transcription service, and the remaining 207 files are manually transcribed by the lead authors to inspect the quality of the data. The details of the concrete quality inspection process can be found in appendix C. The minimum, maximum and mean duration for the audio files collected from LPS are 1.38s, 38.82s and 6.8s, respectively.

[4]Out of 55 unique turkers, four of them did not fill in the post-chat survey.

[5]LPS is accepted based on the speech quality and conversation completeness with HPS.

[6]https://www.speechpad.com/

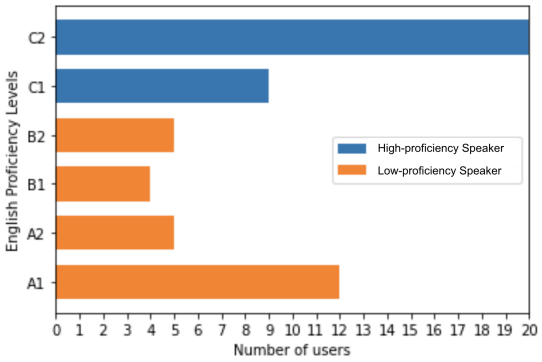


Figure 1: The distribution of CEFR levels in high-proficiency and low-proficiency speakers.

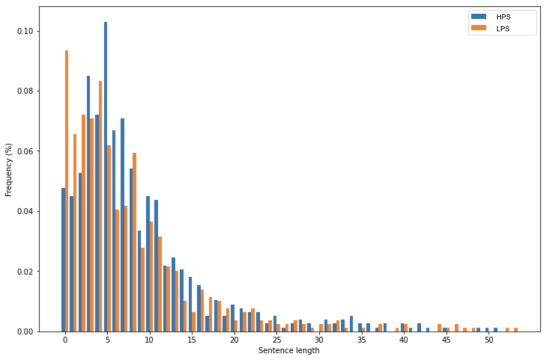


Figure 2: Distribution of number of tokens per turn.

4 GrounDialog Corpus

The primary goal of the data collection was to gather free-form conversations with repair and grounding (R&G) patterns, between high-proficiency (HPS) and low-proficiency (LPS) English speakers. For this work, we constrain ourselves to the domain of job interviews, where an HR hiring manager attempts to schedule an upcoming interview with an interviewee candidate and answers any related questions. We leave the conversations in other domains to our future work.

To analyse the R&G patterns in the collected data from MTurk, we inherit R&G types from previous studies (Dobao and Martínez, 2007; Eszenyi and van der Wijst, 2006; Long, 1983; Foster and Ohta, 2005; Schegloff, 1997; Clark, 1996). The complete list of R&G types is shown in table 2. A detailed explanation of R&G annotation scheme is described below. In addition, similar to other task-oriented dialogue datasets (Budzianowski et al.,

2018; Rastogi et al., 2020), we also annotated the intents and slots for our GrounDialog corpus. To ease our annotation process, we adopted Inception (Klie et al., 2018), which is an open-source annotation software platform.

4.1 Annotation Scheme

4.1.1 Repair and Grounding

R&G can occur over several dialogue turns. It contains the context of the initial communication attempt, questions, and finally a resolution. We tagged these in our dataset as: *Context*, *Question*, *R&G type* and *R&G complete*. The definition for each item type is defined as follows:

- *Context*: the initial utterance as the context of the R&G.
- *Question*: the utterance that triggers the disfluency of the conversation between the two speakers.
- *R&G type*: the R&G type as defined in table 2.
- *Complete*: the utterance that signals the completion of the R&G process.

Note that R&G type is the required item for each R&G annotation, whereas *Context*, *Question* and *Complete* are optional. This is due to the fact that 1) some R&G types can be initiated without the *Context* and *Question* and 2) R&G process maybe not always completed as the conversation moves on.

4.1.2 Intent and Slot

Based on the unified dialog acts ontology defined in He et al. (2022), we proposed ontologies for both intent and slot for our GrounDialog corpus. The full ontology is shown in table 3. The more detailed descriptions for each intent and slot are shown in appendix D.

4.2 Annotation Statistics and Analysis

4.2.1 Repair and Grounding Annotations

The annotations for R&G, Intent and Slot are completed by the lead author. To ensure the quality of the annotations, the lead author and the second author manually inspected each item through comprehensive discussions. The questionable annotation items were corrected if the lead author agreed with the second author.

Figure 3 shows the distribution of different R&G types (a) and R&G related annotations (b) in GrounDialog corpus. There are 269 annotations for R&G types, among which 155 are from HPS and 114 are from LPS. As you can see in figure 3 (a), approximately 30% of the R&G types annotated in HPS utterances are *Proactive Grounding (PG)*. This is due to the fact that the HR manager tends to ask questions that proactively fill in the communication gap and encourage the interviewee candidate to engage in the conversations. For example, in cases when the interviewee candidate forgot to ask questions related to the location of the interview, the HR manager would ask *Do you know how to get to the company?*. On the other hand, as expected, LPS used more *Clarification Request (CR)* in their speech in order to negotiate and confirm critical information for the interview. The example *CR* is shown in table 2.

After including *Context*, *Question* and *R&G complete*, we gathered 604 R&G related annotations, which is nearly 40% of all the dialogue [7]. It can be observed in figure 3 (b) that both HPS and LPS leverage R&G for smoother communication, indicating the potential usefulness of our task setup in terms of negotiation of meaning in natural HPS-LPS conversations.

4.2.2 Intent and Slot Annotation

As for the intent annotations in GrounDialog, there are 1,884 in total, with the number of intents in HPS and LPS being 878 and 1,006, respectively. Figure 4 (left) demonstrates the distribution of intents annotated in the corpus for both HPS and LPS. As you can see, the top two intents are *inform* and *request*, which is similar to larger dialogue datasets like Budzianowski et al. (2018). In our dataset, almost 90% of dialogue utterances have one or two intents indicating the potential of training a language understanding module with our corpus.

Figure 4 (right) presents the distribution of slots annotated in both HPS and LPS responses. There are in total 612 slot annotations, within which 497 slots are annotated from HPS and 115 slots are from LPS. In our GrounDialog corpus, the HPS (i.e. HR managers) tend to give out information in multiple sentences. An example HPS utterance providing concrete location details of the interview is shown below:

[7] Each R&G related annotation is associated with a single utterance. Therefore, the R&G ratio of our dataset is approximately calculated as: 604 / 1569 ≈ 40%.

ID	R&G type	Description	Dialogue Example from GrounDialog
SC	self-correction	When speakers correct own utterances without being prompted to do so by the another person	[Manager]: Are planning to attend the interview? ->Context [Manager]: Are you? ->**SC**
SP	self-paraphrase	A speaker paraphrases the previous response for another speaker to ensure understanding of the response	[Manager]: You have to make a presentation on Webware as company progressing and about its growth ->Context [Interviewee]: I am sorry I did not get that, could you repeat? ->Question [Manager]: You need to tell us your view about present grouth and future growth of company ->**SP** [Interviewee]: Okay. ->Complete
SR	self-repetition	a speaker repeats the previous utterance given the question from the other speaker due to a communication break	[Manager] The interview will be conducted on Monday next week. ->Context [Interviewee] Sorry I did not get the interview time. Could you repeat that? ->Question [Manager] The interview will be conducted on Monday next week. ->**SR** [Interviewee] Got it, thanks. ->Complete
SCL	self-clarification	a speaker provides more information as a supplement to their own previous utterances	[Manager] There will be questions about components. ->Context [Interviewee] Yes, ma'am. ->None [Manager] that you find successful ->**SCL**
QC	question-about -content	a speaker raises question about the contents in the other speaker's response, the contents can include original sentence, phrases, words	[Interviewee] Is there any reimbursement for traveling? ->Context [Manager] reimbursement? ->**QC**
CU	checking-understanding	the manager asks the interviewee a question to check if they understand what the manager has said	[Manager] The interview will be by Monday next week at 11 am. Will you be able to come? ->**CU** [Interviewee] Yes, ma'am. ->Complete
CR	clarification-request	One speaker requests for clarification to get some extra information from the other speaker	[Manager] For the interview there will be 5 of us. ->Context [Interviewee] Could you tell me who exactly will be there during the interview? ->**CR**
TA	tolerate-ambiguity	the manager tolerates the ambiguity in the interviewee's speech and continue the conversation	[Interviewee] Hello. I have some questions about the in-... ->Context [Manager] OK ->**TA**
RH	recheck-history	the interviewee asks the manager questions that refer back to the dialogue history to recheck the information provided in the conversation	[Interviewee] Just to make sure the interview is on next Monday at 4 pm, right? ->**RH** [Manager] yes ->Complete
OH	other-help	the manager senses that the interviewee did not finish the previous sentence so the manager provides "acknowledgement" to help the interviewee continue and complete the unfinished utterance	[Interviewee] Hello. Uh ->Context [Manager] Yes please continue ->**OH** [Interviewee] Uh who will be at the panel? ->Complete
OC	other-correction	the manager finds that the interviewee has made a language mistake and the manager corrects interviewee's mistake	[Interviewee] I want to know is there any green bus meant for traveling? ->Context [Manager] There is no reimbursement. ->**OC**
PG	proactive-grounding	the speaker proactively grounds the information gap	[Manager] Do you know how to get there? ->**PG**

Table 2: A full list of R&G types and their descriptions and dialog examples. The R&G annotations for these examples are also shown for each utterance after '->', and the R&G types are highlighted.

Intent type	inform / request / affirm / small_talk thank_you / hi / self_introduction / bye / reqalts / check_connection negate / welcome / not_sure / select / direct / check_availability / propose sorry
Slots	Location / Interview start time / Interview end time / Day / Duration / Interview attendees / Room number / Transportation

Table 3: Full ontology for intent and slot in GrounDialog.

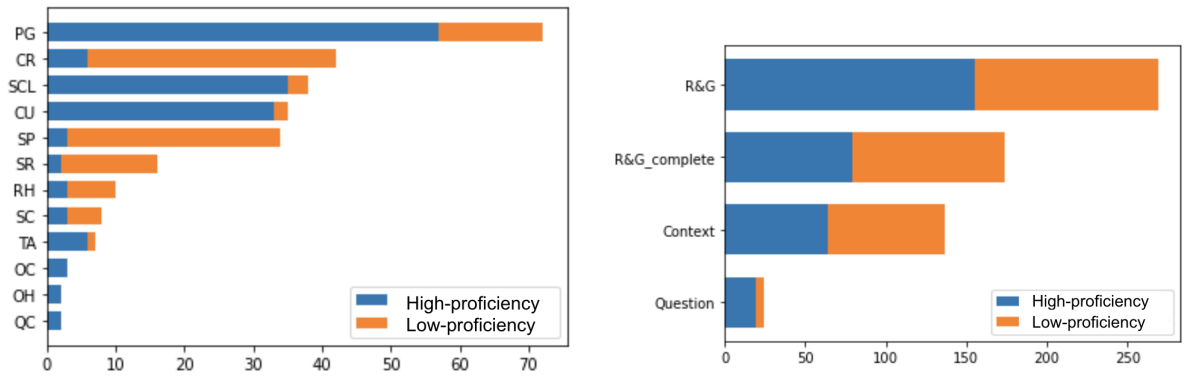
> *Ways to commute to our company: from **Penn Station**; exit via **southwest corner** of the station, walk along the **Broadway** for 3 minutes. The company is on the **right side** of the road.*

In the example, four values for the *Location* slot are in bold. This is also the reason why nearly 45% of the slots in HPS responses are *Location*. In general, HPS produced much more slots compared to LPS, which corresponds to the difference in the number of *inform* intent produced in HPS and LPS responses.

4.3 GrounDialog for Language Learning

As the major focus of this work, it is beneficial to take a deeper look at the R&G related annotations in GrounDialog, and discuss the potential utilities of the dataset for language learning.

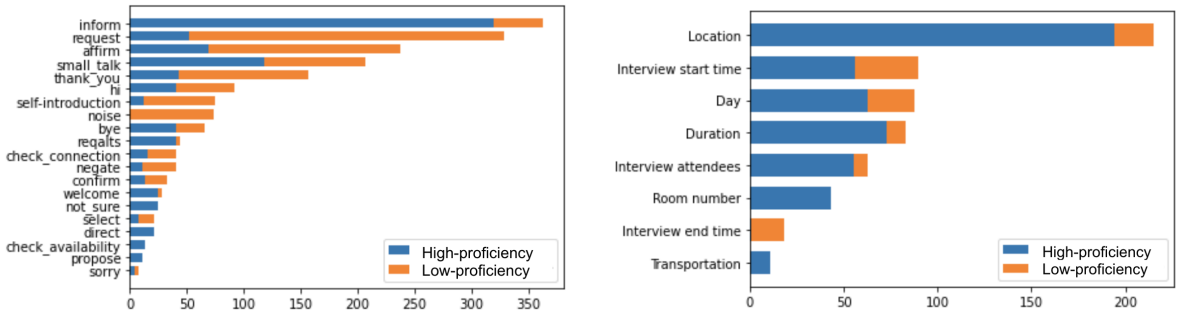
As we have analyzed in the previous section, nearly 40% of the utterances are related to R&G. Figure 5 also presents the distribution of number of R&G annotations per dialogue. Almost 80% of the dialogues have at least four R&G related annotations, showing the richness of R&G patterns in GrounDialog. In general, GrounDialog encapsulates 12 R&G types in the natural HPS-LPS conversations under our task set-up. According to Figure 3(a), the top three R&G strategies for HPS are *proactive grounding (PG)*, *self-clarification (SCL)* and *check understanding (CU)*, whereas LPS mostly uses *clarification request (CR)*, *self-paraphrase (SP)* and *self-repetition (SR)*. This indicates that GrounDialog explicitly encourages LPS



(a) Frequency of R&G types.

(b) Frequency of R&G related annotations.

Figure 3: Frequency of R&G types (left) and R&G related annotations (right) in GrounDialog.



(a) Frequency of intents.

(b) Frequency of slots.

Figure 4: Frequency of intents (left) and slots (right) in GrounDialog corpus.

to request clarification, rephrase or repeat previous utterances in cases when the initial communication with HPS failed.

Besides, we specifically annotated *R&G complete* to mark the sentences that signals the completion of a R&G process. Based on Figure 3(b), among all 269 R&G annotated in GrounDialog, 174 of them are actually completed, leading to a 65% completion rate. Figure 6 shows the distribution of number of *R&G complete* per dialogue. Nearly 80% of dialogues have at least three *R&G complete*, again suggesting the richness of R&G patterns. Also, given the high frequency of R&G related annotations in figure 3(b), we can imply that HPS tends to initiate the R&G much more often compared to LPS in GrounDialog.

From the language learning perspective, learners need R&G patterns to deepen their understanding of the language. For this purpose, GrounDialog can be used to train a chatbot that can generate responses conditioned on our R&G ontology to initiate R&G process, repair the communication gaps, and ground the meanings of conversations for the language learners.

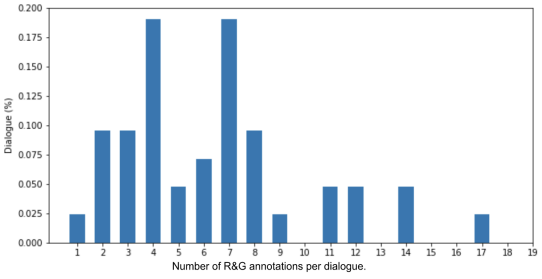


Figure 5: Number of R&G annotations per dialogue.

5 GrounDialog as a Benchmark for R&G in Task-oriented Dialogue

GrounDialog is designed as the first dedicated task-oriented dialogue dataset incorporating R&G patterns in HPS-LPS conversations. To show the potential usefulness of the corpus, we break down the dialogue modelling task into two sub-tasks and report a benchmark result for each of them: R&G detection and dialogue state tracking. Specifically, we performed few-shot learning following recent

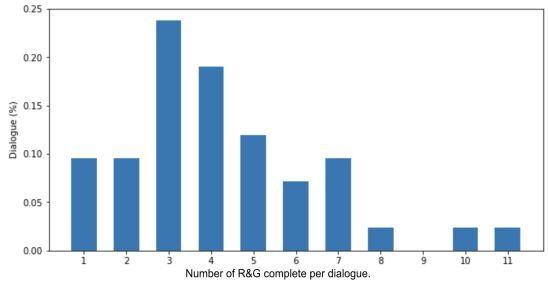


Figure 6: Number of R&G complete per dialogue.

advances in large language models (Brown et al., 2020; Wei et al., 2022), by prompting two most popular large language models, namely ChatGPT and GPT-4 [8], with our carefully engineered prompts for both tasks. The details for each prompt are shown in appendix E.

Model	Slot		Intent	R&G
	Acc	Joint Goal	Acc	Acc
ChatGPT	98.0	88.7	63.1	-
GPT-4	98.2	89.5	65.4	62.1

Table 4: The benchmark results for Dialog State Tracking and R&G detection on GrounDialog.

5.1 R&G Detection

We show that by using the R&G annotations in GrounDialog, an R&G detection model can be trained to determine 1) if communication disfluencies occur; and 2) which type of R&G strategy (as defined in table 2) to choose in order to fix the potential disfluencies incurred in conversations.

Similar to previous section, we prompted GPT-4 for this experiment with the specific prompt defined in appendix E. Note that we tested on 40 out of 42 dialogues, excluding the two we used to design the prompt. For the utterances that do not need R&G, we ask the model to predict "None". The overall detection accuracy is shown in table 4 on the rightmost column [9]. As we can see, prompting GPT-4 can achieve over 62% accuracy on the test dialogues, showing the potential of GrounDialog in training neural models in detecting R&G patterns in natural human-human conversations.

[8]We used `gpt-3.5-turbo` for ChatGPT and `gpt-4` (default 8k version) for GPT-4.

[9]We do not report the results for ChatGPT since it failed to follow the prompt instructions.

5.2 Dialogue State Tracking

A good conversational system requires robust natural language understanding (NLU) and dialogue state tracking (DST) modules. For our benchmark results, we specifically prompted ChatGPT and GPT-4, both of which are popular ground-breaking large language models (LLMs) these days, with our domain-specific prompts. We follow the evaluation metrics for slot extraction in MultiWoz 1.0 (Budzianowski et al., 2018), where overall slot accuracy and joint goal accuracy are reported. For intent classification, we report the general classification accuracy. Table 4 demonstrates the performance of both models in terms of both sub-tasks. As we have only eight slot types in GrounDialog, both models achieved fairly high scores in slot accuracy and joint goal accuracy, with GPT-4 slightly outperforming ChatGPT. With regard to classifying intents, both models achieved over 60% accuracy, even though we have a larger group of intents to classify. These results demonstrate the potential utility of GrounDialog in building a good task-oriented conversational agent with solid NLU and DST modules.

6 Conclusion and Future Work

In this paper, we collected and annotated a new dataset GrounDialog, which is the first dedicated task-oriented dialogue dataset specifically designed for studying repair and grounding in spoken conversations between high-proficiency and low-proficiency speakers. We described the data collection procedure, annotation schemes, and presented a series analysis over the data. In addition, we demonstrated the potential and utility of GrounDialog by performing two tasks: R&G detection and dialogue state tracking. The results showed that GrounDialog can be used to train a conversational agent with the R&G capability. It could be further used to detect communicative gaps, which can be addressed in dialogue design.

In future, we plan to extend GrounDialog to a much larger dataset potentially covering multiple domains other than job interviews. Besides, we will use GrounDialog as a benchmark for a shared task to build task-oriented dialog agent with R&G ability. We will also conduct comprehensive user studies to determine the R&G patterns that are most useful in improving learner's conversational proficiency during language learning. Further, we plan to present findings from the speech data so

researchers can use speech signals along with text to identify repair and grounding related turns.

References

Saul Albert and Jan P de Ruiter. 2018. Repair: the interface between interaction and cognition. *Topics in cognitive science*, 10(2):279–313.

Layla El Asri, Hannes Schulz, Shikhar Sharma, Jeremie Zumer, Justin Harris, Emery Fine, Rahul Mehrotra, and Kaheer Suleman. 2017. Frames: a corpus for adding memory to goal-oriented dialogue systems. *arXiv preprint arXiv:1704.00057*.

Dennis Benner, Edona Elshan, Sofia Schöbel, and Andreas Janson. 2021. What do you mean? a review on recovery strategies to overcome conversational breakdowns of conversational agents. In *International Conference on Information Systems (ICIS)*.

Christina Bennett and Alexander Rudnicky. 2002. The carnegie mellon communicator corpus.

Serge Bibauw, Thomas François, and Piet Desmet. 2019. Discussing with a computer to practice a foreign language: Research synthesis and conceptual framework of dialogue-based call. *Computer Assisted Language Learning*, 32(8):827–877.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz–a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *arXiv preprint arXiv:1810.00278*.

Derek Chen, Howard Chen, Yi Yang, Alex Lin, and Zhou Yu. 2021. Action-based conversations dataset: A corpus for building more in-depth task-oriented dialogue systems. *arXiv preprint arXiv:2104.00783*.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Jiyon Cook. 2015. Negotiation for meaning and feedback among language learners. *Journal of Language Teaching and Research*, 6(2):250.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, Maël Primet, and Joseph Dureau. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *CoRR*, abs/1805.10190.

Modern Languages Division Council for Cultural Cooperation Council of Europe, Education Committee. 2001. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Rahul R Divekar, Jaimie Drozdal, Samuel Chabot, Yalun Zhou, Hui Su, Yue Chen, Houming Zhu, James A Hendler, and Jonas Braasch. 2021. Foreign language acquisition via artificial intelligence and extended reality: design and evaluation. *Computer Assisted Language Learning*, pages 1–29.

Rahul R Divekar, Jaimie Drozdal, Yalun Zhou, Ziyi Song, David Allen, Robert Rouhani, Rui Zhao, Shuyue Zheng, Lilit Balagyozyan, and Hui Su. 2018. Interaction challenges in ai equipped environments built to teach foreign languages through dialogue and task-completion. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 597–609.

Ana M Fernández Dobao and Ignacio M Palacios Martínez. 2007. Negotiating meaning in interaction between english and spanish speakers via communicative strategies. *Atlantis*, pages 87–105.

A Anne Dorathy and SN Mahalakshmi. 2011. Second language acquisition through task-based approach–role-play in english language teaching. *English for Specific Purposes World*, 11(33):1–7.

Réka Eszenyi and Per van der Wijst. 2006. Grounding techniques in computer-mediated classroom tasks. *BELL Belgian Journal of English Language and Literatures*, 4:151–168.

Pauline Foster and Amy Snyder Ohta. 2005. Negotiation for meaning and peer assistance in second language classrooms. *Applied linguistics*, 26(3):402–430.

Wanwei He, Yinpei Dai, Yinhe Zheng, Yuchuan Wu, Zheng Cao, Dermot Liu, Peng Jiang, Min Yang, Fei Huang, Luo Si, et al. 2022. Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10749–10757.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014a. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014b. The third dialog state tracking challenge. In *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 324–329. IEEE.

John F Kelley. 1984. An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2(1):26–41.

Seokhwan Kim, Luis Fernando D'Haro, Rafael E Banchs, Jason D Williams, and Matthew Henderson. 2017. The fourth dialog state tracking challenge. In *Dialogues with Social Robots*, pages 435–449. Springer.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Toby Jia-Jun Li, Jingya Chen, Haijun Xia, Tom M Mitchell, and Brad A Myers. 2020. Multi-modal repairs of conversational breakdowns in task-oriented dialogs. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 1094–1107.

Kai-Hui Liang, Patrick Lange, Yoo Jung Oh, Jingwen Zhang, Yoshimi Fukuoka, and Zhou Yu. 2021. Evaluation of in-person counseling strategies to develop physical activity chatbot for women. *arXiv preprint arXiv:2107.10410*.

Michael H Long. 1983. Native speaker/non-native speaker conversation and the negotiation of comprehensible input1. *Applied linguistics*, 4(2):126–141.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8689–8696.

Emanuel A Schegloff. 1997. Third turn repair. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, pages 31–40.

Veronika Timpe-Laughlin, Tetyana Sydorenko, and Judit Dombi. 2022. Human versus machine: investigating l2 learner output in face-to-face versus fully automated role-plays. *Computer Assisted Language Learning*, pages 1–30.

David R Traum. 1994. A computational theory of grounding in natural language conversation. Technical report, Rochester Univ NY Dept of Computer Science.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Tsung-Hsien Wen, David Vandyke, Nikola Mrksic, Milica Gasic, Lina M Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2016. A network-based end-to-end trainable task-oriented dialogue system. *arXiv preprint arXiv:1604.04562*.

Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL 2013 Conference*, pages 404–413, Metz, France. Association for Computational Linguistics.

A Pre-chat English proficiency self-identification survey

See Figure 7 below.

B Dialogue interface and instructions for High-proficiency and Low-proficiency speakers

See Figure 8 below.

C Audio data quality inspection

This section details the process to inspect the quality of collected audio data. First of all, due to the fact that some collected audio contains long pauses (usually more than 10 seconds without any valid speech), we listened to each audio that is longer than 15 seconds carefully. Then we used `ffmpeg`[10] to truncate the inspected audio which indeed contains long pause to the extend where the audio is natural and continuous. Next, for each audio data, we applied an internal automatic speech recognition tool to detect if the audio is silent all the time. As a result, we discarded all silent audio, and submit the remaining data to SpeechPad [11] for transcriptions.

D Descriptions of Intent and Slot

In this section, we explain different types of intent and slots, and show some examples for better understanding. Specifically, we followed the conventions defined in (He et al., 2022). The descriptions for each intent and slot are shown in Table 5 and 6, respectively.

[10] https://ffmpeg.org/

[11] https://www.speechpad.com/

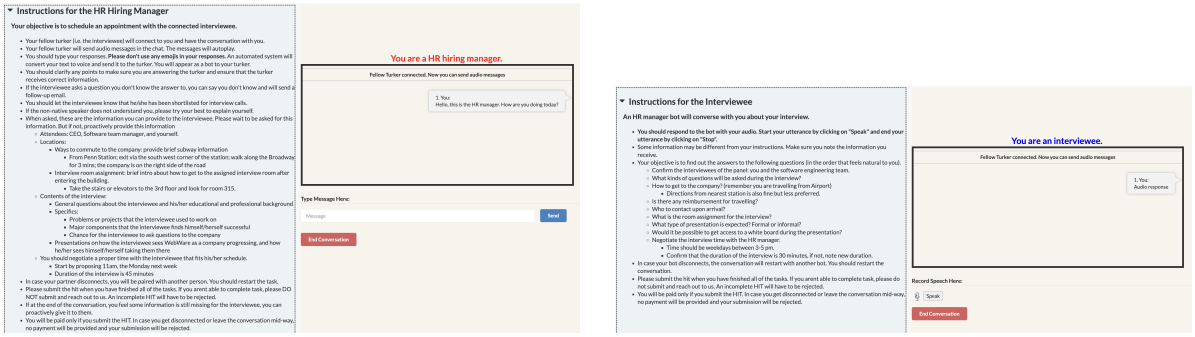
Please fill out the following short survey and submit the information to start the task.

Enter MTurk Worker ID	
Please select your highest level of English conversational proficiency (options displayed low to high proficiency)	○ [*Beginner*] I can interact in a simple way provided the other person is prepared to repeat or rephrase things at a slower rate of speech and help me formulate what I'm trying to say. I can ask and answer simple questions in areas of immediate need or on very familiar topics. AND I can use simple phrases and sentences to describe where I live and people I know.
	○ [*Elementary*] I can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar topics and activities. I can handle very short social exchanges, even though I can't usually understand enough to keep the conversation going myself. AND I can use a series of phrases and sentences to describe in simple terms my family and other people, living conditions, my educational background and my present or most recent job.
	○ [*Intermediate*] I can deal with most situations likely to arise whilst travelling in an area where the language is spoken. I can enter unprepared conversation on topics that are familiar, of personal interest or pertinent to everyday life (e.g. family, hobbies, work, travel and current events). AND I can connect phrases in a simple way in order to describe experiences and events, my dreams, hopes and ambitions. I can briefly give reasons and explanations for opinions and plans. I can narrate a story or relate the plot of a book or film and describe my reactions.
	○ [*Upper Intermediate*] I can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible. I can take an active part in discussion in familiar contexts, accounting for and sustaining my views. AND I can present clear, detailed descriptions on a wide range of subjects related to my field of interest. I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.
	○ [*Advanced*] I can express myself fluently and spontaneously without much obvious searching for expressions. I can use language flexibly and effectively for social and professional purposes. I can formulate ideas and opinions with precision and relate my contribution skillfully to those of other speakers. AND I can present clear, detailed descriptions of complex subjects integrating sub-themes, developing particular points and rounding off with an appropriate conclusion.
	○ [*Proficient*] I can take part effortlessly in any conversation or discussion and have a good familiarity with idiomatic expressions and colloquialisms. I can express myself fluently and convey finer shades of meaning precisely. If I do have a problem I can backtrack and restructure around the difficulty so smoothly that other people are hardly aware of it. AND I can present a clear, smoothly-flowing description or argument in a style appropriate to the context and with an effective logical structure which helps the recipient to notice and remember significant points.
Submit Worker Information	

Figure 7: Pre-chat English proficiency self-identification survey.

E Large Language Models prompts for Dialogue State Tracking and R&G Detection

The prompts we used for experiments in section 5 are shown in Table 7, 8 and 9, respectively. The intent classification and slot extraction task are conducted on a single utterance, whereas R&G detection is conducted on a complete dialogue.



(a) High-proficiency Speaker interface. (b) Low-proficiency Speaker interface.

Figure 8: Dialogue interface and instructions for connected HPS and LPS.

Intent	Descriptions	Example
hi	greeting responses	"Hello" "How are you"
bye	responses for saying goodbye	"Good bye"
thank_you	responses for appreciation	"Thank you"
welcome	denotes a sentence of official texts to welcome	"Welcome and congratulations! You have been shortlisted for the interview"
small_talk	denotes small chats in daily conversation	"So tell me about yourself" "I am fine"
sorry	apologies responses	"I am sorry."
propose	means suggesting to do/offer/recommend something, in order to make the user consider the performance of a certain action, which the manager believes is in the interviewee's interests.	"How about we meet at 11am on next Monday?"
direct	imperative responses that expresses an order	"You need to arrive early for the interview."
request	asking the user about specific attributes (e.g. duration, location)	"What time of the interview suits your schedule?"
select	asking the user to choose a preferred choice from a set of candidates	"Do you want to do it at 11am or 3pm next Monday?"
reqalts	asking the interviewee for more information	"What else information do you want from me?"
affirm	denotes the affirmative responses	"Yes, there is."
not_sure	means the system is not certain about the interviewee's confirmation	"Sorry, I am not sure about this. I will follow up with an email to confirm later"
negate	denotes the negating responses	"No, it is not"
inform	denotes the normal answers to give the information required by the interviewee	"The duration of the interview is 45 mins"
check_ connection	check the connection for the conversation	"Can you hear me?" "There is a lot of background noise"
check_ availability	check the availability of the other person	"Are you able to come?" "Are you okay with the timings?"
confirm	confirm to ground information gap	"Shall we set up the interview?"
self-introduction	introduce personal history and past experiences	"I was a software engineer at another company for 2 years ..."

Table 5: Descriptions and examples for each intent type.

Slot	Descriptions	Example
Interview attendees	The attendees of the interview	The **CEO**, **software manager** and **myself** will be in the interview"
Duration	The duration of an event	"The interview duration is **45 mins**." "The walking duration is **5 mins**."
Room number	The room number of the interview	"Look for room number **315**."
Day	Day of the week	"The interview is on next **Monday**."
Interview start time	The start time of the interview	"Let's aim for the interview next Monday at **3pm**"
Interview end time	The end time of the interview	"The interview will end at **4pm**."
Location	Any location related information	"Please take the subway to **42nd street Time Square**" "You should walk along the **Broadway**."
Transportation	The transportation mentioned by the speaker	"You will have to travel to our office by **train**."

Table 6: Descriptions and examples for each slot type. The slot values are marked in bold.

You need to perform slot extraction tasks. You need to extract "Interview attendees", "Duration", "Room number",
"Day", "Interview start time", "Interview end time", "Location", and "Transportation".

Or if there is no relevant information, you can output "None".
Here are some examples:
[Manager] The CEO, software manager and myself will be in the interview ->"Interview attendees": CEO, "Interview attendees": software manager, "Interview attendees": myself
[Interviewee] How long is the interview? ->None
[Manager] 5 of us will be there in the interview ->None
[Manager] There will be 3 of us in the interview ->None
[Manager] There will be 3 interviewers during the interview ->None
[Manager] The interview will be 45 mins ->"Duration": 45mins
[Manager] The walking duration is 3 mins ->"Duration": 3 mins
[Manager] Look for room number 315 ->"Room number": 315
[Manager] The interview is on next Monday ->"Day": Monday
[Manager] Let's aim for the interview next Monday at 3pm ->"Day": Monday, "Interview start time": 3pm
[Interviewee] Can we have the interview between 3 to 5pm instead? ->"Interview start time": 3, "Interview end time": 5pm
[Manager] We are scheduling the interview for you on next monday at 4pm. ->"Day": monday, "Interview start time": 4pm
[Manager] You will have to travel to our office by train. ->"Transportation": train
[Manager] You can take the elevator to the 3rd floor to find the interview room ->"Location": 3rd floor
[Manager] Please take the subway to 42nd street Time Square ->"Location": 42nd street Time Square
[Manager] way to commute to our company: from Penn station; exit via southwest corner of the station, walk along the broadway for 3 minutes
->"Location": Penn station, "Location": southwest corner of the station, "Location": broadway, "Duration": 3 minutes
[Manager] the company is on the right side of the road ->"Location": right side of the road

Now, let's predict:
[INPUT] ->

Table 7: Prompt for slot extraction. The INPUT tag will be replaced with an actual utterance in the dataset during inference.

You need to perform intent classification tasks. Here are the labels and their definitions:
"hi": greeting responses,
"bye": responses to say goodbye,
"thank_you": responses for appreciation,
"welcome": welcome and tell the interviewees that they have been shortlisted and selected for interview,
"small_talk": small chats in daily conversations,
"sorry": apologies responses,
"propose": means suggesting to do/offer/recommend something, in order to make the interviewee consider the performance of a certain action, which the manager believes is in the interviewee's interests,
"direct": imperative responses that expresses an order,
"select": manager asks the interviewee to choose a preferred choice from a set of candidates,
"reqalts": manager asks the interviewee for more information,
"affirm": denotes the affirmative responses,
"not_sure": means the system is not certain about the interviewee's information,
"negate": denotes the negating responses,
"inform": denotes the normal answers to give the information required by the interviewee,
"check_connection": check the connection for the conversation,
"check_availability": check the availability of the other speaker,
"confirm": confirm to ground information in the chat,
"self-introduction": interviewee introduces personal history and some past working experiences
"request-direction": ask about the direction to the company,
"request-duration": ask about the duration of the interview,
"request-general-info": ask about the general information,
"request-interview-attendees": ask about the interview attendees,
"request-room": ask about the room number of the interview,
"request-time": ask about the timing of the interview,
"request-location": ask about the location of the interview,
Or if there is no relevant information, you can output "None".

Here are some examples:

Hello ->hi
goodbye ->bye
thank you ->thank_you
[Manager] Welcome and congratulations! ->welcome
[Manager] You have been shortlisted for the interview ->welcome
Tell me about yourself ->small_talk
I am fine ->small_talk
I am sorry ->sorry
okay ->affirm
No, it is not ->negate
We don't have any questions. ->negate
Can you hear me? ->check_connection
There is a lot of background noise ->check_connection
Please tell me more about yourself ->request-general-info
Shall we set up the interview? ->confirm
[Manager] You need to arrive early for the interview ->direct
[Manager] What time of the interview suits your schedule? ->request-time
[Interviewee] How long is the interview? ->request-duration
[Interviewee] How to get to the company? ->request-direction
[Interviewee] How can I find the room of the interview? ->request-room
[Interviewee] Please tell me something ->request-general-info
[Interviewee] Where is the interview? ->request-location
[Interviewee] Who will be there for the interview? ->request-interview-attendees
[Manager] What else information do you want from me? ->reqalts
[Interviewee] I have some background in software development ->self-introduction
[Manager] do you want to do it at 11am or 3pm next Monday? ->select
[Manager] Sorry, I am not sure about this ->not_sure
[Manager] Are you able to come? ->check_availability
[Manager] Are you okay with the timings? ->check_availability
[Manager] Do you know how to get there? ->confirm
[Manager] The duration of the interview is 45 mins ->inform
[Manager] How about next Monday at 11am? ->propose
[Manager] The CEO, Software team manager and I will be meeting with you ->inform
[Manager] You should take subway to Penn Station, exit via the south west corner of the station, walk along the Broadway for 3 mins, and the company is on the right side. ->inform
[Manager] You can take the stairs to 3rd floor and search for room 315. ->inform
[Manager] You need to go to the 3rd floor and find the room. ->inform

Now let's predict:
[INPUT] ->

Table 8: Prompt for intent classification. The INPUT tag will be replaced with an actual utterance in the dataset during inference.

You should extract repair and grounding patterns in the conversations. Here are the labels and their definitions:
"Context": the initial utterance as the context of the repair and grounding pattern,
"Question": the utterance that triggers the disfluencies of the conversation between the two speakers,
"self-paraphrase": a speaker paraphrases the question for another speaker to ensure understanding of the question,
"checking-understanding": the manager asks the interviewee a question to check if they understand what the manager has said,
"clarification-request": request for clarification to get some extra information,
"other-correction": the manager finds that the interviewee has made a language mistake and the manager corrects interviewee's mistake,
"other-help": the manager senses that the interviewee did not finish the previous sentence so the manager provides "acknowledgement" to help the interviewee continue and complete the unfinished utterance,
"question-about-content": a speaker raises question about the contents in the other speaker's response, the contents can include original sentence, phrases, words,
"recheck-history": the interviewee asks the manager questions that refer back to the dialogue history to recheck the information provided in the conversation,
"self-clarification": a speaker provides more information as a supplement to their own previous utterances,
"tolerate-ambiguity": the manager tolerates the ambiguity in the interviewee's speech and continue the conversation,
"proactive-grounding": the speaker proactively grounds the information gap that is not about duration,
"self-correction": when speakers correct their own utterances without being prompted to do so by another person,
"self-repetition": a speaker repeats the previous utterance given the question from the other speaker due to communication break,
"Complete": the utterance that signals the completion of the repair and grounding process and it is normally responding to affirmative questions.
Or if there is no repair and grounding pattern, you can output "None"

Here are two examples for the task: please provide annotation for each utterance below after '->'
Dialogue #1:
[Interviewee] Hallo. ->None
[Manager] Hi, how are you? ->None
[Interviewee] I am fine ->None
[Manager] Good. ->None
[Manager] Shall we set up the interview? ->proactive_grounding
[Interviewee] Yes ->Complete
[Manager] What do you know about the company's product/services? ->None
[Manager] What time are you free tomorrow ->self-correction
[Interviewee] I can only do it from 3 to 5pm. ->None
[Manager] I see. In that case, do you want to do it at 3pm? ->checking-understanding
[Interviewee] Yes, I can. ->Complete
[Manager] Do you know how to get here? ->proactive_grounding
[Interviewee] Yes ->Complete
[Manager] Okay. ->None
[Manager] there will be questions about components ->None
[Interviewee] Yes, ma'am. ->None
[Manager] that you find successful ->self-clarification
[Manager] The CEO, Software team manager and I will be meeting with you. ->None
[Interviewee] Okay. ->None
[Manager] I am sorry, Can you repeat your last response? ->Question
[Interviewee] I said okay. ->self_paraphrase
[Manager] Thank you. ->Complete
[Interviewee] I need to know how to get to the company ->None
[Manager] Are you traveling from the airport or train station? ->clarification-request
[Interviewee] The airport ->Complete
[Manager] Do you have any questions? ->proactive_grounding
[Interviewee] I want to know is there any green bus meant for traveling? ->None
[Manager] There is no reimbursement. ->other-correction
[Manager] Then we will see you Monday at 11. ->None
[Interviewee] Good bye. ->None
[Manager] bye ->None

Dialogue #2:
[Interviewee] Hello. Uh ->Context
[Manager] yes please continue ->other-help
[Manager] Hello I am your hiring manager ->None
[Interviwee] Hello ->None
[Manager] I wanted to inform you that you have been shortlisted for an interview ->Context
[Manager] which will be next week on friday ->self-clarification
[Manager] What does your wife do? ->Context
[Interviewee] My wife? ->question-about-content
[Manager] Yes ->Complete
[Interviewee] Are we going to have the interview? ->proactive_grounding
[Manager] Yes good morning ->Complete
[Interviewee] How long is the interview? ->None
[Manager] The duration of the interview will be 45 minutes. ->Context
[Interviewee] Sorry, I did not catch that. What did you say? ->Question
[Manager] the duration of the interview will be 45 minutes. ->self-repetition
[Interviewee] Oh okay. ->Complete
[Manager] When is your flight? ->Context
[Interviewee] Sorry, what did you ask? ->Question
[Manager] At what time will you be leaving for the flight? ->self-paraphrase
[Interviewee] On Monday 2pm ->Complete
[Manager] do you have any questions? ->proactive_grounding
[Interviewee] How to get to the company? ->None
[Manager] ways to commute to our company: from Penn station; exit via southwest corner of the station, walk along the broadway for 3 minutes. ->None
[Manager] the company is on the right side of the road. ->None
[Interviewee] Okay ->None
[Interviewee] How can I find the room of the interview? ->None
[Manager] you will enter the building and look for room 315 on third floor ->None
[Interviewee] Okay great. ->None
[Manager] good luck for the interview. Have a great day, bye. ->None
[Interviewee] Thank you so much. ->None
[Interviewee] Just to make sure the interview is on next Monday at 4pm, right? ->recheck-history
[Manager] Yes ->Complete
[Interviewee] Okay awesome. Thank you ->None
[Manager] No. Thank you ->None
[Interviewee] Bye. ->None

Now, please give the prediction for the new conversation. Forget the history and do not generate new dialogue.

[INPUT DIALOGUES]

Table 9: Prompt for R&G detection. The INPUT DIALOGUES tag will be replaced with a complete dialogue during inference.