Building Representations of Common Ground with Intelligent Agents: Papers from the 2011 AAAI Fall Symposium (FS-11-02)

Towards Overcoming Miscommunication in Situated Dialogue by Asking Questions

Matthew Marge and **Alexander I. Rudnicky**
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213 USA

Abstract

Situated dialogue is prominent in the robot navigation task, where a human gives route instructions (i.e., a sequence of navigation commands) to an agent. We propose an approach for situated dialogue agents whereby they use strategies such as asking questions to repair or recover from unclear instructions, namely those that an agent misunderstands or considers ambiguous. Most immediately in this work we study examples from existing human-human dialogue corpora and relate them to our proposed approach.

Introduction

Ambiguity and miscommunication occur regularly in conversations and decrease the amount of information that agents can correctly understand. These problems prevent agents from adding concepts under discussion to the common ground they have with users. Agents can overcome this limitation by detecting ambiguities as they occur in conversation, then asking a user appropriate questions to clarify meaning. There are several challenges associated with this process. The agent must first detect that there is a communication problem (e.g., in situated dialogue with a user, the agent may resolve that a move the user requested is impossible or that it had low confidence in understanding what the user said). Secondly, the agent must classify the *type* of the problem. Having done so the agent should select a recovery strategy and ask questions that can resolve the communication problem. After receiving feedback, the agent is able to fix the problem. The agent can then add the concepts under discussion to the common ground.

We propose to develop an approach that allows agents to formulate appropriately diagnostic questions and to use responses to repair unclear instructions. Miscommunication occurrences such as ambiguity and misunderstanding happen often in situated dialogue (among other domains), especially with route instructions (i.e., navigational directions). For this reason, we propose to study this problem in the navigation domain. Moreover, route instructions are sequential; agents can generate a direct correspondence between language and action and place actions in the correct order. Agents can amend or update route instructions by dividing them into parts based on subgoals. Route instructions also have clear measures of success (e.g., whether or not an agent reaches a goal or subgoal; the extent to which an agent deviates from the optimal path). Communication about navigation has been studied previously, and as a result corpora are available and are a good source of data for present purposes.

Background

Miscommunication in Dialogue

Traditionally miscommunication encompasses misunderstanding, non-understanding, and ambiguity in dialogue. Hirst and colleagues (Hirst et al. 1994) defined differences between misunderstanding and non-understanding in human-human dialogue. *Misunderstanding* refers to when a dialogue partner interprets an utterance in a way that is not in line with what the speaker intended. Note that this dialogue partner still believes that communication occurred to some extent. However, *non-understanding* refers to the event where the dialogue partner fails to interpret an utterance at all. Non-understandings are noticed right away in human-human dialogues, while it may take additional turns for a dialogue partner to detect a misunderstanding. When a misunderstanding occurs, one dialogue partner may initiate a repair, whereby the partner tries to find and amend an invalid assumption. When a non-understanding occurs, one partner may initiate a recovery by trying to get the conversation back on track (Skantze 2007). *Ambiguity* occurs when a speaker provides insufficient information to determine which specific referent among a set of candidates is the intended one. Experimental evidence suggests that ambiguity occurs often in human-robot situated dialogue (Liu, Walker, and Chai 2010).

Error Detection and Recovery in Spoken Dialogue Systems

Many researchers have studied approaches to detecting and recovering from errors in spoken dialogue systems. Bohus and Rudnicky studied several approaches to combining evidence from the speech recognizer, parser, and dialogue manager to estimate confidence about an input utterance (Bohus and Rudnicky 2002). Walker and colleagues evaluated many features of the natural language understanding and di-

alogue management components of a telephony-based dialogue system to detect errors (Walker, Wright, and Langkilde 2000). Skantze has also investigated methods to overcoming errors in spoken dialogue that exist in human-human communication (Skantze 2007).

Identifying Unclear Instructions

Identifying if an interpreted instruction is ambiguous or not correctly understood is the first step towards overcoming miscommunication between people and dialogue agents. In situated dialogue, awareness and reference to the environment can help identify and recover from ambiguous instructions. We propose a component that assesses interpretation confidence using evidence from the environment and from the content of a speaker's commands.

We can do this by examining dialogue corpora that relate to navigation for factors that lead to unclear instructions (e.g., the HCRC Map Task (Anderson et al. 1991) and SCARE (Stoia et al. 2008) dialogue corpora). Here are possible sources of information:

- The *confidence scores from an automatic speech recognizer* (trained on the appropriate vocabulary for the domain) indicates uncertainty in understanding, more specifically recognition, of a person's speech. The agent's speech recognizer, when applied to interpreting conversational speech, will be valuable for determining non-understanding.
- The *likelihood of a parse* of recognized speech can help resolve miscommunication by measuring misunderstanding if a parse is unlikely, or non-understanding if no parse of recognized speech is possible. This is also the first line of confidence assessment in typed dialogue interaction.
- Once the agent interprets an instruction, it may generate several action plans and select one to execute. The degree of *plan ambiguity* (Carberry 2001) the agent generates after interpreting an instruction can also determine unclear instructions, particularly with respect to actions in the environment. For example, if a person instructs the agent to 'move to the door', the agent may need to choose among several doors. The agent can repair this ambiguity by clarifying which door the person intended.
- The *inconsistency between the plan and mapping of actions to the environment* (i.e., a requested move is impossible) is an indicator of situated understanding confidence. For example, if an instruction would require an agent to move through a wall, the agent should inform the speaker that the requested move is not possible and ask to amend the plan.
- The amount of ambiguous language present can help us detect unclear instructions. We propose that the agent can learn to detect ambiguous language by analyzing word co-occurrence statistics in navigation dialogues of the direction giver directly before the follower ask questions about ambiguity (e.g., *which*-questions).
- Reliable prosodic/acoustic features in the speech of instructions that suggest the instruction was unclear can be a source of misunderstanding. For example, f0 and duration in the speech signal may vary when the instruction giver hesitates.

Existing spoken dialogue architectures such as the Olympus Spoken Dialogue Framework (Bohus et al. 2007) often include a component devoted to assigning a confidence score to an interpreted utterance. Olympus uses the Helios confidence annotator to combine evidence from the speech recognizer, parser, and dialogue manager. The system uses this evidence to decide which interpretation of an input utterance most closely matches the speaker's intention. The dialogue system then executes the interpretation with the highest confidence score.

Corpora Examined

We will discuss examples of miscommunication occurrences in two human-human dialogue corpora, the HCRC Map Task Corpus and the SCARE corpus.

HCRC Map Task Corpus

The HCRC Map Task corpus consists of 128 unscripted English human-human dialogues (64 participants) in the navigation domain (Anderson et al. 1991). Participants were tasked with verbally discussing how to replicate a path on one participant's (complete) map on the partner's (incomplete) map. The experimenters assigned participants to either the role of the *direction giver* or the *direction follower*. Both dialogue partners had a 2-dimensional schematic map of a fictional area, containing landmarks in various places. The key variation between the dialogue partners was that the direction giver had a labeled path on his map from start to finish, while the direction follower only had the start point labeled. The follower needed to trace the path that the direction giver described on his own map. They could not see each other's maps.

Since the participant task in this corpus is navigation, there are clear measures of communicative success. For subgoals, one measure of success is if the direction follower successfully reached the landmark or landmarks that the direction giver described. The overall task was a success if the direction follower arrived at the finish point on the direction giver's map.

This corpus fits our interests because recovery strategies for resolving miscommunication are very common in this corpus; in fact, the Map Task's dialogue act coding scheme contains labels for such strategies (Carletta et al. 1997). The *check* code for a turn indicates when a dialogue partner needs to confirm something with the other that he is not certain about. The *query-yn* and *query-w* codes label questions one dialogue partner asks of the other (yes/no questions or otherwise). These codes can be used to identify recovery strategies as they occur in the dialogues.

SCARE Corpus

The SCARE corpus is a set of 15 unscripted English human-human dialogues in which one partner directs another through a virtual world (Stoia et al. 2008). The roles of the dialogue partners are similar to the Map Task, with a

direction giver instructing a direction follower through manipulation tasks in the environment. Unlike the Map Task, the giver had a 2-dimensional schematic map and the follower was situated directly in the virtual world. The giver had a first-person perspective of what the follower was seeing (they were not sitting at the same computer). Thus the dialogue partners shared knowledge not only by spoken dialogue but also by monitoring the position and gaze of the follower in the corresponding virtual environment.

In addition to navigation-style instructions, this corpus contains instructions for manipulating objects in a virtual environment. Since miscommunication can occur in dialogue about both navigation and manipulation, they can help inform the design of language agents that implement recovery strategies. Partners communicate by dialogue and through monitoring the video feed of the follower. The corpus permits active tracking of both forms of input since it contains audio and video recordings of all sessions.

Recovery Strategies

There are three types of recovery strategies that we will investigate that may allow the agent to determine when to ask questions and the type of questions to ask. We relate these strategies to examples in the Map Task and SCARE human-human dialogue corpora.

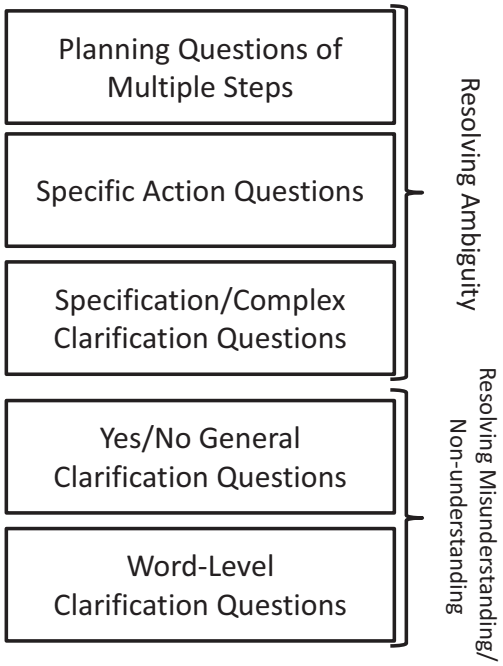

Figure 1: Taxonomy of the questions to be employed in repair/recovery strategies.

Resolving Ambiguity

Recovery strategies to resolve ambiguity occur when there are multiple possibilities for converting a language instruction to an action sequence. The agent will need to make several decisions to resolve ambiguity. First, the agent must determine the 'depth' level to ask a question (see the taxonomy in Figure 1). Second, the agent needs to determine its human dialogue partner's capabilities (i.e., Can the person answer questions about the environment reliably? Does the person possess better or worse knowledge of the environment than the agent?). Third, the agent may need to resolve an ambiguous perspective (i.e., Was the person referring to the agent's perspective or his own perspective?). The types of questions that will resolve ambiguity will be planning questions, specific action questions, and specification questions:

- Planning questions of multiple steps:

  Direction Giver (DG): we are going to go (noi) (sil) due south (sil) straight south (sil) and (noi) (sil) then we're going to g– (sil) turn straight back round and head north (sil) past an old mill (sil) on the right (sil) hand side
  Direction Follower (DF): **due south and then back up again**
  (Map Task corpus)

  In this case the direction giver gave a very long route instruction to the follower. The follower interpreted a portion of the instruction, but verified the most immediate part of the plan to start the task. The follower was uncertain if he correctly interpreted a part of the plan, but was somewhat uncertain of his interpretation of the instruction. To repair this issue, the follower asked a question regarding a part of the giver's plan that he heard.

- Specific action questions (e.g., a bargein when one person detects an ambiguity about a high-level goal):

  DG: well there is a button that controls it but
  DF: (sil)
  DG: oh
  DF: **controls what**
  (SCARE corpus)

  In this example, a button in the virtual environment controls a cabinet, but the follower is unsure what the button controls. The follower knows it controls something in the environment, and asks the giver about it.

- Specification/complex clarification questions:

  DG: (sil) um (sil) ok there's a door in the room go through it (sil) (pause)
  (sil) and then sort of (sil) do the u thing (sil) (sil) go to the right (pause)
  (sil) yeah there'll be another door there (sil) (pause)
  DF: **(sil) this one or the**
  DG: um
  DF: **left (sil)**
  (SCARE corpus)

  This is an example of a specification clarification question that is also situated in the environment. In this case the follower arrives at a point where there are multiple doors, and must determine which door the giver was referring to. The follower gave the giver only two possible answers to respond.

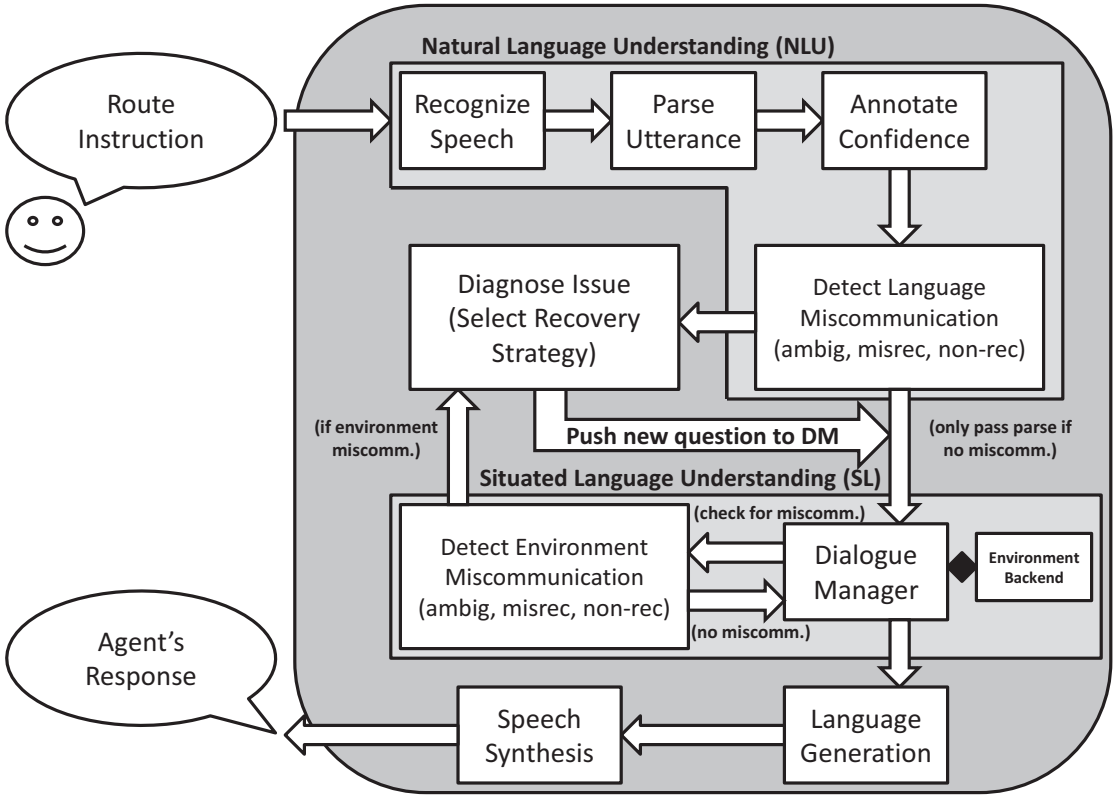

Figure 2: A situated dialogue system pipeline with natural language understanding, situated language understanding, and recovery components.

Resolving Misunderstanding and Non-understanding

Recovery strategies to resolve misunderstanding and non-understanding occur when there is high confidence that the speech recognizer interpreted many words in an utterance incorrectly. They may also occur when a parse of an input utterance is unlikely. The questions that the agent can ask here reference information at a lower level than that used when resolving ambiguity. Specifically the two types of questions the agent will use are yes/no clarification questions and word-level clarification questions:

- Yes/no general clarification questions:

  DF: **have i to pass by the savannah (noi)**
  DG: no
  DG: you've to come down
  (Map Task corpus)

  Yes/no clarification questions are restricted to questions that have only a 'yes' or 'no' answer. Here is an example from the Map Task corpus where the follower asked if moving around a specific landmark (i.e., the savannah) was part of the giver's map.

- Word-level clarification questions:

  DG: until you (sil) you get over the top of the slate mountain
  DF: **over the top of the (noi)**
  DG: slate mountain
  DF: don't have a slate mountain
  (Map Task corpus)

  Word-level clarification questions repair misrecognized or non-recognized utterances. In this example, the follower may not have heard the giver correctly. To recover from this, the follower repeats back a portion of the utterance that he correctly recognized in order to request the part he wanted the giver to repeat.

Resolving Plan Execution

Recovery strategies during plan execution occur when the user interrupts the agent as it executes its current action plan. This could be because the user decides on another action plan or the agent is not following the user's intentions. The agent must determine when the user can bargein when it is executing an instruction. In this case, the agent should ask questions specifically for repairing or revising existing instructions.

The agent may also resolve a plan execution when a requested action is not possible. This requires the agent to plan out an instruction using the access it has to the environment to determine if an obstacle or the environment properties prevent a move from completing. When the user requests an impossible instruction, the agent should inform the user of the problem and prompt for an amended instruction (this may be requesting an entirely new instruction, or amending a part of the initially requested instruction).

Proposed Understanding Components

Figure 2 presents the current design for the situated dialogue agent pipeline with recovery strategies. We examine the case where a user gives a situated dialogue agent (e.g., a mobile robot) a route instruction. The utterance first enters the natural language understanding (NLU) component of the system. Here, the component recognizes the input speech and assigns a confidence score to the recognition. The transcribed text is passed along to a parser, which assesses the likelihood of the parse. This confidence annotator for the NLU combines the speech recognizer's confidence with the parser's confidence into one score. A miscommunication detector (language only) then analyzes the transcribed, parsed utterance for any evidence of miscommunication if the confidence score is below a learned threshold.

If the confidence score is low, the detector sends all available information related to the utterance for diagnosis of the problem and to determine a recovery strategy. The selected recovery strategy is pushed to the dialogue manager (DM) as a new question. The DM serves as the decision engine for the dialogue system, and in this case sends the question along the traditional pipeline (natural language generation and speech synthesis) to the user.

If the system detects no miscommunication at this point, the parsed instruction is sent to the Situated Language Understanding (SL) component. First, the DM attempts to execute the parsed route instruction in the environment. The DM sends the input utterance and environment information to a miscommunication detector (language and environment). If there is a problem with executing the route instruction (e.g., it is not possible) or is ambiguous with respect to references in the environment, all relevant information is sent for diagnosis, and the system selects a recovery strategy. The system then pushes a new question addressing the recovery strategy to the DM, which is sent along to the user to address. If no miscommunication is detected in either the NLU or SL components, the agent executes the route instruction in the environment and updates the user on its status.

To conclude this section, we provide examples of interactions we expect the system to handle, using the question types in Figure 1.

- Planning questions of multiple steps:

  User: Turn right, go down the hallway, and search the three rooms on the left.
  System: There is only one door in that hallway. Should I still go down it and search? Or did you mean something else?

- Specific action questions:

  U: Turn right, go down the hallway, and search the three rooms on the left.
  S: I'm down the hallway. Did you tell me to search the three rooms on the *right*? Or was it the *left*?

- Specification/complex clarification questions:

  U: Go to the door on the right.
  S: Did you mean the door on *your* right or *my* right?

- Yes/no general clarification questions:

  U: Go down the hallway and search the three rooms on the right.
  S: You said to go down the hallway and search the three rooms on the right. Did I get that right?

- Word-level clarification questions:

  U: Go to the black tile.
  S: Did you say to move to the *blue* tile?

Future Work

Corpus Analysis Procedure

A set of well-annotated human-human navigation corpora is available for analysis (e.g., the HCRC Map Task (Anderson et al. 1991) and SCARE (Stoia et al. 2008) corpora discussed earlier). The coding schemes of these corpora can be used to build a distribution of question types. The agent may learn from the recovery strategies of instruction followers in the tasks associated with these corpora. These data can help us understand how often followers ask questions by building tabulations of all the questions asked using existing corpus annotations. At the same time we can determine the most common types of questions that followers ask.

After designing a set of recovery strategies, we intend to develop a probabilistic model that uses speech features, language features, and environment features to determine when an agent should ask questions given (1) the last exchange between the speaker and the agent, (2) the current state of the world, and (3) the last n steps of the dialogue. The accuracy of this model can be evaluated using annotations from existing corpora, though environment information may require new data collection. One challenge we anticipate here is the reliability of the existing corpora for finding useful, consistent examples of recovery.

Evaluation

The agent's performance at detecting unclear instructions can first be assessed on labeled corpora. This will require evaluating the confidence annotator as it combines evidence from the input speech, the transcribed language, and the agent's environment. Although several annotated corpora

exist, we may in addition consider collecting data specifically for this experiment. We also intend to integrate this confidence annotator into an existing human-robot spoken dialogue system, TeamTalk, for realtime evaluation (Rudnicky et al. 2010). Doing so will allow us to determine how well recovery works in practice and to capture additional information, such as the success of iterative attempts at repair. We expect to analyze measures of task completion and subjective questionnaires on interaction with the updated system.

Summary

Situated dialogue between dialogue partners can contain ambiguity and other forms of miscommunication that make reliable information exchange difficult. For language agents to robustly work in situated dialogue domains, they must be able to detect these occurrences during interaction with humans. We presented a method for language agents to recover from miscommunication in dialogue by asking questions. The proposed method will combine evidence from the speech signal, natural language understanding, and environment components to detect occurrences of ambiguity, misunderstanding, and non-understanding. Specifically, we examined situated dialogue in the navigation domain, where one dialogue partner gives route instructions to another. We showed the types of questions such an agent might ask and discuss examples from existing human-human dialogue corpora.

References

Anderson, A. H.; Bader, M.; Bard, E. G.; Boyle, E.; Doherty, G.; Garrod, S.; Isard, S.; Kowtko, J.; McAllister, J.; Miller, J.; Sotillo, C.; Thompson, H. S.; and Weinert, R. 1991. The hcrc map task corpus. *Language and Speech* 34(4):351–366.

Bohus, D., and Rudnicky, A. I. 2002. Integrating multiple knowledge sources for utterance-level confidence annotation in the cmu communicator spoken dialog system. *Technical Report CS-190*.

Bohus, D.; Raux, A.; Harris, T. K.; Eskenazi, M.; and Rudnicky, A. I. 2007. Olympus: an open-source framework for conversational spoken language interface research. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, NAACL-HLT-Dialog '07, 32–39.

Carberry, S. 2001. Techniques for plan recognition. *User Modeling and User-Adapted Interaction* 11:31–48.

Carletta, J.; Isard, S.; Doherty-Sneddon, G.; Isard, A.; Kowtko, J. C.; and Anderson, A. H. 1997. The reliability of a dialogue structure coding scheme. *Comput. Linguist.* 23:13–31.

Hirst, G.; McRoy, S.; Heeman, P.; Edmonds, P.; and Horton, D. 1994. Repairing conversational misunderstandings and non-understandings. *Speech Communication* 15(3-4):213 – 229. Special issue on Spoken dialogue.

Liu, C.; Walker, J.; and Chai, J. 2010. Ambiguities in spatial language understanding in situated human robot dialogue. In *Proceedings of the AAAI Fall Symposium on Dialog with Robots*.

Rudnicky, A.; Pappu, A.; Li, P.; Marge, M.; and Frisch, B. 2010. Instruction taking in the teamtalk system. In *Proceedings of the AAAI Fall Symposium on Dialog with Robots*.

Skantze, G. 2007. Error handling in spoken dialogue systems: Managing uncertainty, grounding and miscommunication. Ph.D. diss., Dept. of Speech, Music, and Hearing, KTH, Stockholm, Sweden.

Stoia, L.; Shockley, D. M.; Byron, D. K.; and Fosler-Lussier, E. 2008. Scare: A situated corpus with annotated referring expressions. In *Proceedings of LREC '08*.

Walker, M.; Wright, J.; and Langkilde, I. 2000. Using natural language processing and discourse features to identify understanding errors in a spoken dialogue system. In *Proceedings of the 17th International Conference on Machine Learning*, 1111–1118.