The TeamTalk Corpus: Route Instructions in Open Spaces

Matthew R. Marge
Carnegie Mellon University, mmarge@cmu.edu

Alexander I. Rudnicky
Carnegie Mellon University, ar28@andrew.cmu.edu

Follow this and additional works at: http://repository.cmu.edu/lti

Part of the Computer Sciences Commons

Published In
Proceedings of Workshop on Grounding Human-Robot Dialog for Spatial Tasks.

This Conference Proceeding is brought to you for free and open access by the School of Computer Science at Research Showcase @ CMU. It has been accepted for inclusion in Language Technologies Institute by an authorized administrator of Research Showcase @ CMU. For more information, please contact research-showcase@andrew.cmu.edu.
The TeamTalk Corpus: Route Instructions in Open Spaces

Matthew Marge and Alexander I. Rudnicky
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
Email: {mrmarge, air}@cs.cmu.edu

Abstract—This paper describes the TeamTalk corpus, a new corpus of route instructions consisting of directions given to a robot. Participants provided instructions to a robot that needed to move to a marked location. The environment contained two robots and a symbolic destination marker, all within an open space. The corpus contains the collected speech, speech transcriptions, stimuli, and logs of all participant interactions from the experiment. Route instruction transcriptions are divided into steps and annotated as either metric-based or landmark-based instructions. This corpus captured variability in directions for robots represented in 2-dimensional schematic, 3-dimensional virtual, and natural environments, all in the context of open space navigation.

I. INTRODUCTION

There is general agreement that navigation is essential for mobile robots to accomplish tasks. Spoken language interaction is one way to move robots, by communicating about space and distance via route instructions. This paper describes a corpus that we believe will help the human-robot dialogue community better understand the nature of spoken language route instructions when they are directed toward robots. The corpus contains over 1600 route instructions from a total of thirty-five participants, all fluent speakers of English (some were non-native English speakers).

Although dialogue will be a necessary aspect of understanding people’s language interaction with robots, we must first capture their initial intentions when they give robots directions. We collected a corpus for this specific purpose: Given a static scenario with a robot capable of understanding natural language, what will people say to navigate the robot to a designated location? All directions required the robot to only move within an open space, and not around a more complex environment. There was only one landmark: another visible robot near the designated location. This environment setup allowed us to systematically vary the configurations of the robots in the scenarios and observe how people adjusted their route instructions to the changes.

Direction giving has been of enduring interest to the natural language processing and robotics communities; many groups have collected and released similar corpora to the TeamTalk corpus. Perhaps the best known is the Map Task corpus, a collection of dialogues that has one person (a giver) provide directions to another person (a follower) to move along a prescribed path on a shared map [1]. The SCARE [10], GIVE [5], and CReST [4] corpora are dialogue corpora with a similar navigation task, but present environments in different ways (for SCARE and GIVE the environment was a virtual world; for CReST the environment was a computer-generated schematic overhead view). The IBL corpus [6] captured people’s spoken language directions to a static robot in a real-world miniature model of a city. The corpus collected for building the MARCO [7] route instruction interpreter captured people’s typed instructions for navigating another person around a virtual environment. The current corpus consists of spoken language directions within the confines of open space navigation, systematically varying location and robot orientation. None of the corpora above apply systematic variation to stimuli to elicit directions. In addition, our corpus only captures what the giver of directions says (similar to the IBL corpus but unlike the others), since the follower is described as an artificial agent.

II. OPEN SPACE NAVIGATION TASK

In the open space navigation task, participants provided verbal instructions that would let a mobile robot move to a specified location (we present details from a previous paper about this work below [8]). Participants viewed a static scene on a computer monitor with two robots, Mok and Aki, and a destination marker (Mok was the actor in all scenarios). The experimenter told participants to give instructions as if they were observing the environment but not situated in it. In other words, the robots could hear participants but not see them (so participants could not mention themselves in their instructions).

The experimenter indicated that Mok was capable of understanding natural language. Participants were free to assume they could use naturally occurring language when they provided instructions to Mok (i.e., no formally structured language was necessary). The experimenter showed participants the orientations of the robots and told them they could use these orientations in their instructions. The robots did not move in the scenes, so participants did not see how the
robot responded to their instructions (participants received no feedback on their instructions).

All scenarios required participants to give all the necessary steps to move Mok to the goal destination in a single recording. Participants accomplished this by using a recording interface that did not disrupt their view of the robots’ environment. They spoke their instructions through a close-talking headset microphone. The recording interface allowed participants to playback their instructions and re-record them if necessary. The experimenter also told participants to think about their instructions before recording; this was meant to deter participants from replanning their instructions as they spoke. We time-stamped and logged all interaction with the recording interface, and provide it as part of this corpus.

We varied the orientations of the two robots, Mok and Aki, and the location of the destination marker for the recordings. The robots were each in one of four orientations, directly pointing forward, right, left, or backward. The destination marker was in one of four locations, directly in front of, behind, to the left of, or to the right of Aki. Participants viewed environment representations in 2-dimensional schematics, 3-dimensional virtual scenes, or real-world scenes. For the schematic and virtual environments, we varied the three factors using a full-factorial design, resulting in 64 configurations (randomly ordered) of the robots and destination marker per session. Each participant for these recordings provided 64 instructions. Participants each gave 8 instructions when viewing real-world scenes (destination varied four ways, Mok’s orientation varied two ways). We describe these environments in detail below.

A. Schematic Environment

A segment of participants viewed 2-dimensional schematic representations of the two robots and a destination marker. The schematics presented a birds-eye view of the scene, with the robots presented as arrows and the destination marker symbolized by a purple circle (See Figure 2(a)). The arrows indicated the perspective of the robots. There was no sense of scale in the environments; participants could not use metric information when instructing the robot.

B. Virtual Environment

Some participants gave instructions to a robot situated in a 3-dimensional virtual environment. We developed the environment using a virtual map builder and USARSim, a robot and environment simulator that uses the Unreal Tournament game engine [3]. The environment contained two Pioneer P2AT robots and a transparent purple destination marker. The environment was set to a standing eye-level view, at a height of about 1.8 meters (see Figure 2(b)). Walls in the space were far enough away from the robots that participants did not use them as landmarks. The recording setup used two monitors, one to show a full-screen representation of the environment, and one for the recording interface.

Half of the participants in this study were told that the two robots were seven feet apart (i.e., the distance condition); the experimenter did not specify any distance to the remaining participants (i.e., the no-distance condition). The distance condition allowed participants to provide instructions to the robot using metric information. The environment did not provide any other indication of scale.

C. Natural Environment

The environment for this group of participants was similar to the virtual condition, but participants gave instructions in-person (see Figure 2(c)). The robots were represented as bins in the space; eyes on the top of the bins indicated each robot’s orientation. Participants were standing in a gymnasium, not sitting at a computer. No logging information was collected, only verbal recordings.

D. Participation

Thirty-five self-reported fluent English speakers participated (ten viewed the schematic environment, fourteen viewed the virtual environment, eleven viewed the real-world scene environment). There were twenty-two male and thirteen female participants, ranging in age from 19 to 61 ($M = 28.4, S.D. = 9.9$). We recruited participants by posting to the Carnegie Mellon Center for Behavioral Decision Research webpage [4]. Participants earned $10.00 for completing the task.

\[ \text{http://www.cbdr.cmu.edu/experiments/} \]
TABLE I
CORPUS INFORMATION ABOUT THE TYPES OF ROUTE INSTRUCTIONS THAT PARTICIPANTS GAVE IN THEIR RECORDINGS.

<table>
<thead>
<tr>
<th>Environment</th>
<th># Participants</th>
<th># Route Instructions</th>
<th>Transcription Method</th>
<th>Disfluency Annotations?</th>
<th>Step Annotations?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schematic</td>
<td>10</td>
<td>640</td>
<td>In-house</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Virtual (distance)</td>
<td>7</td>
<td>448</td>
<td>6 Crowdsourced, 1 In-house</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Virtual (no-distance)</td>
<td>7</td>
<td>445</td>
<td>6 Crowdsourced, 1 In-house</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Real-world</td>
<td>11</td>
<td>86</td>
<td>In-house</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>

III. DATA COLLECTION

The corpus contains 1,619 verbal route instructions and is summarized in Table I. On average there are 23.1 words per instruction (S.D. = 18.2). Total word count was 37,442 with 751 unique terms (some terms are words; others are annotations). There are 640 route instructions directed to robots in the schematic environment (sixty-four recordings from each of ten participants). These instructions were transcribed in-house according to the CMU Communicator transcription conventions [2]. The corpus also contains logs of participants’ interactions with the recording interface. A past study used these logs to derive the amount of time participants took to formulate their instructions after viewing a scene (i.e., “thinking time”) [8].

We collected more route instructions from the virtual environment scenarios (893 instructions, 64 from each of fourteen participants). This segment of the corpus also evenly divides instructions into those collected from participants that had an absolute sense of scale in the environment (labeled in the corpus as “distance” instructions) and those that did not (labeled in the corpus as “no-distance” instructions). Workers from Amazon Mechanical Turk transcribed recordings from twelve participants using the same guidelines as before (see [9] for a description of the approach). The real-world study yielded 86 route instructions², all of which were transcribed in-house.

²Three virtual trials and two real-world trials are missing due to a recording interface issue with two participants.

IV. CORPUS ANNOTATION

As the corpus was used for a study of direction giving [8], the route instruction transcriptions were annotated for discrete actions per instruction (i.e., ‘steps’) and disfluencies. We describe the annotation coding scheme below.

A. Step Annotation

We divided participant recordings into discrete steps. In this corpus, we defined a step as a sequence of words that represented any single action that would move Mok (the acting robot) to a subgoal. The following example has two steps, divided by a forward slash ‘/’:

Mok turn left / and move forward three feet

The first step in this instruction is a rotation, while the second moves Mok to the goal destination.

We annotated steps as one of two types, absolute steps and relative steps. An absolute step is a step that explicitly includes a measured distance or turn rotation. This includes simple turns (e.g., “turn right”) that we assume to be rotations of ninety degrees. Both of the steps above are absolute steps. Discretized measures that aren’t metric distances are also absolute steps (i.e., “move forward three steps”). Absolute step examples:

move right four feet
move forward about seven feet
turn left ninety degrees
turn around

A relative step is one that mentions a landmark as a subgoal or reference point. In this study, the only possible landmark was Aki, the static robot in all scenarios. Below are some examples of relative steps:

go forward until you’re in front of Aki
go forward half the distance between you and Aki

A single annotator divided transcribed recordings into steps and labeled them as absolute or relative. Marge and Rudnicky [8] discuss the proportions of absolute and relative steps for the schematic and virtual scene subsets of the corpus. 58.9% of the steps in schematic scene segment of the corpus had absolute steps, while the remaining 41.1% were relative steps. We divided step analysis for the virtual scene instructions into...
distance and no-distance segments. The no-distance instructions had a similar breakdown to the schematic instructions; 68.5% of the steps were absolute and 31.5% were relative. The distance instructions were markedly different in nature; most were absolute (93.5% compared to 6.5% relative). The study that used this corpus found that when aware of distance, participants included it in nearly all instructions.

B. Disfluency Annotation

The schematic instructions in this corpus were transcribed in-house; the transcriptions include annotations for disfluencies. The three types of disfluencies we annotated were fillers (e.g., uh or um), mispronunciations (i.e., when the speaker does not pronounce an intended word) and false starts (i.e., when the speaker speaks then abruptly begins again).

The in-house transcriber annotated fillers by surrounding fillers by forward slashes, as shown below:

I think /uh/ you should turn left

Mispronunciation annotations have the uttered word fragment and what the transcriber believed to be the intended word. Square brackets surround the entire mispronunciation annotation and within the brackets the parenthesized text has the intended word, as follows:

move [sev (seven)] feet forward

False starts indicate all words in an uttered phrase that were abruptly abandoned for a new phrase. Angled brackets surround the abandoned phrase, like below:

<turn right> no turn left

Often a mispronunciation and false start occur in the same uttered speech, these occurrences include both annotations. See below for an example:

<turn [ri (right)]> no turn left

The crowdsourced transcriptions did not have these annotations because the crowdsourced transcriptions were found to be word-accurate. Results pertaining to their accuracy correctly marking disfluencies is forthcoming.

C. Session Log Information

The experiment recording software logged all interface activity. The logs recorded four or more items per trial: presentation of a stimulus then starting, stopping, replaying, and accepting a recording (we define trial here as one of the sixty-four recordings per participant). The log information can be used as an indicator of cognitive load for participants as they formulated instructions. More specifically, the elapsed time between when participants viewed a scene and pressed the ‘Record’ button can measure load on participants as they formulated instructions. When ‘thinking time’ was long, we hypothesize that the participant incurred a high cognitive load while formulating an instruction for that scene.

V. CONCLUSION

This paper described the TeamTalk corpus, a new corpus that contains the speech and transcriptions of route instructions directed to robots. Fluent speakers of English gave verbal directions to a robot in an open space that would allow the robot to move to a specified location. The corpus captures speakers’ intentions when giving navigational directions and provides useful information for researchers studying spatial language and route instructions. The most immediate impact of this corpus will be to help build grammars for human-robot dialogue systems and for general language analysis (i.e., vocabulary building and language modeling, in combination with other resources). The corpus can be found at http://www.cs.cmu.edu/~robotnavcps.

ACKNOWLEDGMENTS

This work was sponsored by the Boeing Company and a National Science Foundation Graduate Research Fellowship.

REFERENCES