Journal Title
XX(X):1–4
©The Author(s) 2016
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Placeholder title for *Natural Language Instructions for Collaborative Manipulation*

Rosario Scalise*, Shen Li*, Henny Admoni, Stephanie Rosenthal, and Siddhartha S. Srinivasa

Abstract

This paper presents a dataset of natural language instructions for object specification in manipulation scenarios. It is comprised of 1607 individual written instructions which were collected via online crowdsourcing. This dataset is particularly useful for researchers who work in natural language processing, human-robot interaction, and robotic table-top manipulation. It provides a benchmark of image/instruction pairs to be used in system evaluations as well as uncovers inherent challenges in table-top object specification. Example code is provided for easy access via Python.

Keywords

Class file, $\LaTeX\,2_\varepsilon$, *SAGE Publications*

Introduction

In this paper, we provide a corpus of natural language instructions used to instruct a robot in a collaborative table-top manipuation setting. It is common for cluttered table-tops to elicit object reference ambiguity in tasks requiring natural language communication. Because understanding and generating natural language is essential for human-robot collaboration, we propose a dataset which provides scenarios accompanied by human generated instructions describing table-top objects subject to specification ambiguity. Figure ??RS: change to harder scene? shows an example of one such scenario. These scenarios reveal the challenges inherent in specification ambiguity such as object uniqueness, object proximity, and perspective. Finding methods to resolve these ambiguities are of particular interest to researchers in the feilds of robotics and natural language processing. We include the images used to elicit the instructions and data collected in evaluating instruction preformance as well as lexical categorization data. Instructions were generated by human participants via Amazon's Mechanical Turk (AMT) [cite]. Data coders labeled each instruction with the number of objects it could refer to without any inference, and the type(s) of perspective present in each instruction. We also provide an evaluation dataset collected via AMT in which the original set of instructions was measured through correct object resolution accuracy by a seperate set of participants.

Image annotation datasets have seen significant growth in popularity in the past RS: number years, particularly within the computer vision community. Two such examples are ImageNet Deng et al. (2009) and VQA Antol et al. (2015). These datasets which combine images and natural language have proved invaluable as baselines for informing and evaluating algorithms relating language, semantics, and object recognition. Our dataset aims to provide an analogous, albeit more domain-specific set of baseline scenarios.

In a similar vein, robotics researchers have put considerable efforts into establishing corpora for use in navigational instructions which are inherently spatial tasks. In Skubic et al. (2012), the authors collect a set of indoor route following instructions by showing participants a target object within a virtual room and asking them to instruct an avatar to go to the target. They examine varying the particpants' prompts, the type of agent being instructed, and the perspectives they should use. Similarly, MacMahon and Stankiewicz (2006) collects first-person navigational instructions in a virtual 3D environment, and then evaluates these instructions with a seperate set of participants to test instruction effectiveness. They place emphasis on environmental landmarks rather than perspective or ambiguity. We note that while these works focus on navigational trajectories, we focus on defining manipulation goals that are not necessarily constrained to particular trajectories.

RS: potentially add something citing Howard Towards the goal of fluent interfacing for robotic table-top manipulation, Bisk et al. (2016) / Bisk et al. (2016) contribute dataset results. The authors generated configurations of blocks on a tabletop where each block is uniquiely identified by a number or symbol on its faces. They asked participants to give instructions to transform one configuration to another and recorded the instructions given as series of steps called Problem-Solution Sequences.

The dataset we present draws from a closely related line of work, with a greater emphasis on the ambiguities often encountered in real-world settings when issuing manipulation instructions. For example, in manipulating objects amongst clutter, it is common to inadvertently give descriptions which can refer to multiple targets. This raises

[1] Sunrise Setting Ltd, UK
[2] SAGE Publications Ltd, UK

Corresponding author:
Alistair Smith, Sunrise Setting Ltd Torbay Innovation Centre, Vantage Point, Long Road, Paignton, Devon, TQ4 7EJ, UK.
Email: alistair.smith@sunrise-setting.co.uk

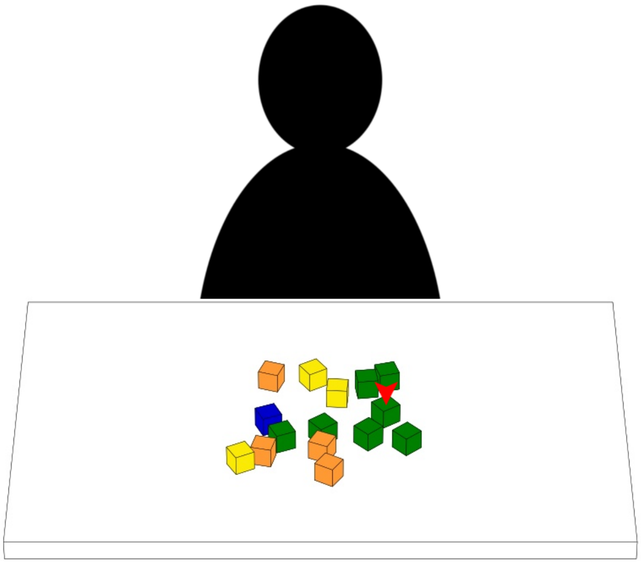

Figure 1. Image 6, Configuration Version 1

the issues of resolving spatial references, understanding visual feature landmarks, and utilizing perspective in an effective manner. Li et al. (2016) uses this dataset to present initial results that explore these issues.

Researchers interested in utilizing natural language commands to perform manipulation tasks (particularly in a cluttered environments where human-robot collaboration will occur) should find the presented corpus valuable in establishing a baseline of unconstrained human tendencies when describing ambiguous objects on a table-like surface. NLP researchers will also find this a good source for language which involves instructions as well as a high frequency of spatial references. Applications range from providing robotic accessability in the home to people with mobility disabilities, enabling mutual understanding via language in a robot-human collaboration task such as a robotic sous chef in a resturaunt kitchen, and interfacing with an industrial PCB pick-and-place robot where discrete components can differ only slightly in appearance.

The following could be included in their own sections if necessary:

- Include where to find dataset (website?)
- Include License (if any)

MISC - Might have Acknowledgements / Funding section

Data Collection Methods

Participants and Demographics

RS: double check participant numbers All participants were sourced through Amazon's Mechanical Turk [footnote]. 120 participants were involved in the first study and 356 participants were involved in the second study. These two groups of participants were mutually exclusive.

Demographics were collected for each user. We asked participants to report their Age, Gender, Occupation, Computer Usage, Hand Dominance, English as Primary Language, and Experience with Robots, RC Cars, First-Person-Shooter (FPS) Video Games, and Real-Time-Strategy (RTS) Video Games.

We took care to ensure study designs avoided any confounding participant specific variables such as color-blindness. For example, red blocks were not used because red-green color blindness is the significantly predominent form of the deficiency. **?**

Study 1: Collecting Natural Language Instruction Corpus

The purpose of the first study was to collect a corpus of instructions. In particular, instructions elucidated from scenarios where often times blocks are not easily uniquely identified and spatial references or other visually apparent traits must be utilized to aid in object identification. We created a set of stimulus images from which a randomly selected subset would be presented to each participant. An example of a stimulus image is shown in .

Each image in the set consists of 15 randomly-spaced, randomly-colored blocks (orange, yellow, green, or blue) reseting on a table with a silhoutted figure behind the table. The silhouetted figure represents a partner the participant is to interact with. We synthesized 14 images and selected 2 blocks from each image which are indicated by an arrow to elicit a range of possible participant responses. In total, there are 28 unique scenarios in the set of stimuli.

We presented each participant with 14 randomly selected stimuli from the full set of 28. For each stimulus, we asked the participant to instruct the silhoutted figure to pick up the block indicated by the arrow. We randomly selected and told each particpant the partner they were instructing was either a human or a robot. We held this assigment constant across all stimuli shown to a particular participant. [footnote: the silhoutted figure was not changed] The participant entered their response into a textbox in typed natural language. We also asked the particpant to subjectively rate the difficulty of creating the instruction for each stimulus on the Likert scale (1 (easy) to 5 (difficult)). At the end of the sequence of stimuli, we asked each participant 1) if they employed any particular strategy in completing the task, 2) how challenging they found the overall task, and 3) for their general comments.

For each instruction, we captured meta-data such as the total time duration it took the participant to write the instruction after being shown the stimulus. We collected 1607 instructions in total.

RS: Should we discuss the 'down-sampled' set of 1400 with approx. 50 instructions per scenario? Should we describe the issues we ran into such as sentences that do not parse

Study 2: Evaluating Corpus

Given the original set of stimuli and the set of instructions collected in the first study, we collected responses from new participants indicating which block the participant believed the accompanying instruction specified within the stimulus scenario. For each stimulus, we removed the indication arrow.

We showed the participant the stimulus/instruction pair with an interactive selection interface RS: figref which displayed a circle around each block as the particpant hovered their mouse. When the participant clicked on the block they believed the instruction specified, the circle would change to a checkered pattern and remain overlayed on the image. The participant was free to change their selection as many times as necessary and their time to select was unconstrained. We gave the participant no feedback to indicate the accuracy of their selection.

For a subset of 1400 instructions randomly sampled from the full set after removing invalid responses RS: this could be explained at end of Study1 if necessary, we collected the number of blocks a participant clicked on while selecting their final answer as well as the final answer itself. We collected approximately RS: check this 10 final answers per stimuls/instruction pair across all particpants and computed the accuracy for a particular instruction.

RS: am I missing anything else we collected here? What about other general questions we may have asked?

Data Coding

We coded the corpus of instructions for two distinct purposes. The first purpose was to determine the ambiguity present within an instruction. To address this, we defined a set of criteria based upon whether an explicit perspective was used, and if so which perspective the instruction was given in according to the set: { partner, participant, neither }. We labeled each instruction with one of these perspective or the ambiguous 'unknown' perspective. In addition, we labeled each instruction according to how many blocks could possibly be specified without using any inference. As an example, if there were 3 blocks fitting the description 'red block near you', then the block ambiguity for this instruction would be 3.

The second purpose was to develop a categorization of descriptive features. We performed hand-coded word binning for each unique word in the set of all instructions according to a set of categories we deteremined during analysis.

Perspective and block ambiguity was manually coded by four raters. In order to establish interrater reliability, we required each of the four coders to code the same 10% of the full dataset. We corrected for any discrepincies in coding until a high interrater reliability was achieved. This was confirmed by conducting pairwise Cohen's κ tests and averaging the results.For coding perspective, the κ value was .85 which indicates a very high interrater reliability. For coding block ambiguity, the κ value was .68 which indicates a high interrater reliability. Once the reliabilities were established, each of the four coders processed one quarter of the data and the results were merged.

Dataset Details

Instruction Corpus Table

The file named NLInstructions.csv which contains the initially collected corpus of natural language instructions is formatted so that each line corresponds to an individual instruction. Note that there are multiple instructions which correspond to a given image. Here, we briefly describe each field's relationship to an instruction.

The *Instruction* field contains the string(s) entered by the participant along with an *Index* for ease of reference. Each instruction corresponds to a particular *Scenario* image which is named by the originating image and its variation (1 or 2). The *AgentType* indicates whether the instruction was given to a human or robot.

Each instruction was also given a *Difficulty* rating by the participant which indicated how challenging the they thought a particular prompt was when they generated the instruction. Additionally, the *TimeToComplete* a given instruction was recorded. The duration was started immediately following the current prompt loading and ended when the participant clicked next after entering their instruction and difficulty rating for the prompt.

At the end of the study, an individual participant is asked a series of questions to collect additional information and demographics. Note the participant is only asked once, however, for each instruction the corresponding answers are repeated in the table for ease of reference. These questions include *Strategy* which asks the user to enter their general comments on the strategies they might have used to generate instructions throughout the study, *Challenging* which asks how challenging the participant found the overall study, *GeneralComments* which asks for the participant's general comments on the study, the participant's *Age*, *Gender*, and *Occupation*, the *ComputerUsage* habits of the participant indicated by the number of hours per week, whether the participant has *EnglishAsFirst* language, and finally the participant's experiences with robots, RC cars, first-person-shooter video-games, and real-time-strategy video-games indicated by a number between 1-5 according to the Likert scale (where 1 is strongly disagree and 5 is strongly agree). Additionally, we include an assigned unique user ID to each participant that is distinct from the Amazon Mechanical Turk ID. A summary of all fields is shown in Table 1.

- In addition to outlining the components of the dataset, this section can also include salient findings/nice statistics/plots.
- Example: If we set 70% correct block selection accuracy as the benchmark, we observe about 30% of instructions do not meet this criteria and would not constitute an instruction which can be executed reliably.
- Fields shown in 1

Accessing the Dataset

- Describing accessor functions.

References

Deng J, Dong W, Socher R, Li L.J., Li K, and Li F (2009) ImageNet: A lrage-scale hierarchical image database. In:*Computer Vision and Pattern Recognition, 2009*.

Antol A, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick C.L., and Parikh D. (2015) VQA: Visual Quesion Answering In:*International Conference on Computer Vision, 2015*.

Table 1. Overview of fields in Natural Language Instruction Corpus.

Parameter	Note	Data
`Instruction`	Given by participant as response to prompt	Sentence(s) in String format
`Index`	Corresponds to individual instruction	Number from 1-1607
`Scenario`	Configuration by image and variation	String with img. num. 1-14 and v. 1-2
`AgentType`	Agent participant is instructing	String from {robot, human}
`Difficulty`	Rating for a scenario given by participant	Number from 1-5 (5 being most difficult)
`TimeToComplete`	Duration participant took to create instruction	Time in hours:minutes:seconds
`Strategy`	Comments on strategy for used	Sentence(s) in String format
`Challenging`	Comments on overall study difficulty	Sentence(s) in String format
`GeneralComments`	Comments on general on overall study	Sentence(s) in String format
`Age`	Age of participant	Number indicating age in years
`Gender`	Gender of participant	String
`Occupation`	Occupation of participant	String
`ComputerUsage`	Hrs/wk participant uses a computer	
`DominantHand`	Dominant hand of participant	String from {Right, Left}
`EnglishAsFirst`	Participant's first language	Number from {1 = Yes, 0=No}
`ExpWithRobots`	Particpant exp. with robots	Number from 1-5 on Likert scale
`ExpWithRCCars`	Participant exp. with RC cars	Number from 1-5 on Likert scale
`ExpWithFPS`	Exp. with first-person-shooter video games	Number from 1-5 on Likert scale
`ExpWithRTS`	Exp. with real-time-strategy video games	Number from 1-5 on Likert scale
`ExpWithRobotComments`	Comments on exp. with robots	Sentence(s) in String format
`InternalUserID`	Particpant ID unaffiliated with Amazon ID	Number from 1-600

Skubic, M, Alexenko, T, Huo, Z, Carlson, L, and Miller, J (2012). Investigating spatial language for robot fetch commands. In:*Workshops at AAAI Conference on Artificial Intelligence*.

acMahon, M, and Stankiewicz, B (2006). Human and automated indoor route instruction following. *Def*, 2(6), 4.

Bisk Y, Yuret D and Marcu D (2016) Natural Language Communication with Robots. In:*Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics*. San Diego

Bisk Y, Marcu D and Wong, W (2016) Towards a Dataset for Human Computer Communication via Grounded Language Acquisition. In:*Proceedings of the AAAI 2016 Workshop on Symbiotic Cognitive Systems*. Phoenix.

Li, S, Scalise, R, Admoni, H, Rosenthal, S, and Srinivasa, S (2016). Spatial References and Perspective in Natural Language Instructions for Collaborative Manipulation. In:*Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN).*