Natural Spatial Description Generation for Human-Robot Interaction in Indoor Environments

Zhiyu Huo
Dept. of Electrical and Computer Engineering
University of Missouri-Columbia
Columbia, MO, USA
zhiyuhuo@mail.missouri.edu

Marjorie Skubic
Dept. of Electrical and Computer Engineering
University of Missouri-Columbia
Columbia, MO, USA
SkubicM@missouri.edu

Abstract—**This paper proposes a spatial language generation system to communicate with a person about the location of an object in an indoor environment. It aims at finding a short, accurate and human-like description for building a natural and friendly interface between robots and humans using spatial language interaction. The system performs an inverse procedure to spatial language grounding which links natural commands to robot actions. The system works in two steps. It will first search for the best matching grounding model which describes the spatial relations between the target object and the references; then it will generate the natural language by mimicking a human's talking style. A corpus of 149 spatial language commands for an indoor environment fetch task is used to train the language generation model. An early-stage experiment is conducted and the results illustrate a potential for further development.**

Keywords—spatial language, robotics, language generation

I. INTRODUCTION

Robots have begun to leave the laboratory and are constantly increasing in the office, homes, and living apartments for the elderly. The interest in how a robot can be of assistance in our daily life continues to grow. A survey shows that one of the top tasks wanted by elderly users is to let robots help them with household tasks such as fetching objects [1]. The elderly also prefer to use natural language rather than learning a computer programmer's approach to interacting with robots. For most users untrained in computer languages, the meaning of spatial language interaction is to send a natural language command to a robot and to get a response in natural language as well. This led to a call for more study of robot understanding and on generating a more user-friendly spatial language for the robot. In our research project, two complimentary challenges in human-robot spatial language interactions are proposed: (1) To give robots commands based on the non-scientific community's need for a natural spatial language, and (2) to develop robots that can inform (speak to) humans by communicating in natural spatial descriptions. A corpus of human spatial descriptions has been collected and a spatial language-driven robot system for fetching has been developed. The majority of the first challenge has been met; this paper presents a methodology to solve the second challenge.

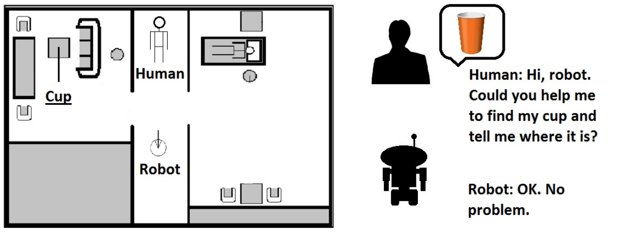

Cup
Human
Robot
Human: Hi, robot. Could you help me to find my cup and tell me where it is?
Robot: OK. No problem.

Fig.1 The scenario of an object searching and language generation task performed by a robot in a home environment

Fig.1 is a scenario of an object searching and language generation task performed by a robot in a home environment. The human user is standing in the hallway between the living room and the bedroom, and he cannot remember where he put his cup. He wants the robot to find it and tell him the location of it so that when he needs it, he can easily go right to it. In this scenario, the human user expects the robot to give a description like "*The cup is on the table behind the couch in the living room*" or "*Walk into the living room, then turn right and move forward, you will see the cup on the table,*" which is a natural and friendly way of assisting and providing enough information to assure successful retrieval. Here, we focus on the generation of static spatial language which has been introduced in [2]. Static spatial language uses objects as references to describe a target location, i.e., "*behind the couch*" or "*on the table next to the bed*". Our previous work [3] showed that older adults tended to use furniture items as reference objects.

In a language generation task, the robot needs to construct a model of its working environment by using the sensory information it collected from the environment to generate the static spatial language description. Moreover, such a description may be long and may have a complex structure which will make it difficult to be generated by a language template. These make it a more challenging task. However, this kind of human-like spatial language provides more intuitive navigation information for a human user, particularly an elderly user. To enable the robot to give easily understood spatial descriptions to a human user, we designed a multi-step

978-1-5090-0898-8/16/$31.00 ©2016 IEEE

system which first models the spatial relations from the sensory information on the working environment, and then generates natural language from this intermediate result.

II. Methodology

A. Overview

To interact with users via human-like spatial language, robots should have the capability to use understandable human language to give spatial descriptions based on the working environment. For our project, we would like the robot to give a natural language description of the position of a target object. For the example shown in Fig.2(a), the expected corresponding description is "*The cup is on the table in front of the couch in the living room*". In such a task, we let W denote the working environment, and p denote the location and the orientation of the human user. From the definition of an objective function $h(\varphi,p,W)$ of the natural description φ, the robot will search for a spatial description φ' with the largest function value:

$$\varphi' = argmax_{\varphi} h(\varphi, p, W) \quad (1)$$

The objective function determines the policies to select a spatial description which should: a) have accurate information for the human user to reach the target object, b) match the human spatial language syntax and human's habit on language and c) use the fewest number of words. However, to directly train the cost function by samples of φ, p and W is a problem of great complexity. Here, we propose a multi-step process that splits the workflow into three steps:

(1) Model the Environment: the robot will build an environmental model which includes all the detected objects in its working environment until it finds the target object (Fig.2(b)). All the objects in the environment are recorded (Fig.2(c)). The information about an object is described by an Entity model which includes a category name, a coordinate vector, an orientation value, a 3D grid voxel model and a unique ID of the object.

(2) Generate a Grounding Model: The system uses an objective function to generate a grounding model. In our work, the grounding model is a standard representation for the spatial relations between the entities in the environment. In this system, the reference-direction-target (RDT) format presented in our previous work [2] is used for the grounding model. This model could represent a grounding by multiple spatial relations between two objects. The grounding model of a command can be represented as a chain of RDT nodes, which describes a sequential action list or reference-based description allowing the robot to move to find the target object (Fig.2(d)). The RDT generation here is a reverse procedure of spatial command grounding in [2].

(3) Generate Natural Language: After getting the RDT grounding model, the system generates natural language by a model learning using our spatial language corpus (Fig.2(e)) [4].

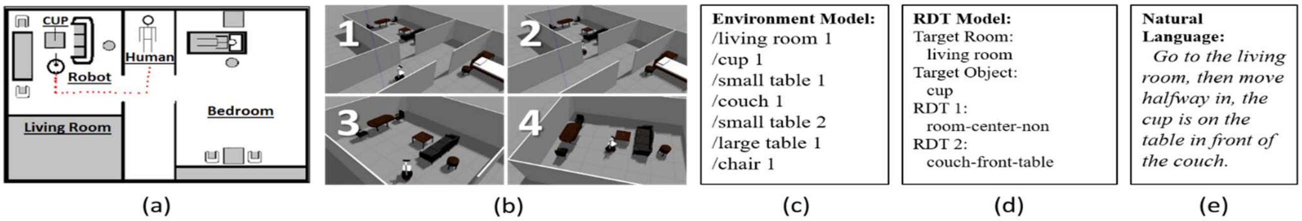
B. Build Environment Model

The first step to generate a static spatial language description is building an environment model. The robot builds prior internal knowledge about the objects in the working environment so that it can recognize the objects and build their entity models. The robot will keep building the environment model during the object searching process in the working space until it gets the target object. Thus it can finally get a set of N entities $\varepsilon=\{e_1,...,e_N\}$ in the working environment. The list ε does not only include the entities of a room such as wall, furniture and the target object, but also has a robot entity for the last pose of the robot. Our system uses a depth camera as the robot sensor, which is used to recognize the category of an object and also capture its geometric features such as size, shape and orientation.

C. Spatial Groundings in the Map

Next, the robot generates a static RDT grounding model, which includes several RDT nodes. The entities list ε is used to build a spatial relation list $\Gamma(\varepsilon)=\{\gamma_1,...,\gamma_M\}$. The list Γ includes M combinations between any two entities. For each combination, we use $\gamma_m=\{F_{direction}(e_1,e_2),F_{distance}(e_1,e_2)\}$ to represent two histogram vectors of direction and distance as the features of a spatial relationship. This is called the world state feature (WSF), which describes the spatial relations in the environment. After building the WSF, we calculate the probability $P_{\Gamma}(y)$ of each possible RDT node y which will be used later in the objective function.

To seek the best solution over all RDT nodes, an objective function is proposed. Let $\{y_1,...,y_K\}$ denote K RDT nodes that can be extracted from the environment. A weight w_k is 1 if the RDT node y_k is selected to generate the spatial language description and is 0 if not selected. A number v_{k1k2} is a value in [0,1] which is the probability of the two RDT nodes y_{k1} and



Environment Model:
/living room 1
/cup 1
/small table 1
/couch 1
/small table 2
/large table 1
/chair 1

RDT Model:
Target Room:
living room
Target Object:
cup
RDT 1:
room-center-non
RDT 2:
couch-front-table

Natural Language:
Go to the living room, then move halfway in, the cup is on the table in front of the couch.

Fig.2 The steps to generate a natural spatial description by robot.

y_{k2} appearing together. We let $P_{yk}=P_{\Gamma}(y_k)$ which is the probability of y_k in the environment. Then we can compose the following objective function for the combination of all the K RDT nodes which is:

$$O(W) = \frac{\left(\sum_{k=1}^{K} w_k P_{y_k} + \sum_{k1=1}^{K}\sum_{k2=1}^{K} v_{k1k2} P_{y_{k1}} P_{y_{k2}}\right)}{\sum_{k=1}^{K} w_k + \sum_{k1=1}^{K}\sum_{k2=1}^{K} v_{k1k2}} + l$$

$$W = [w_1, \dots w_K] \qquad l = \frac{\alpha}{\sum_{k=1}^{K} w_k} \qquad (2)$$

Here W is a vector of all the w_k values. To get the best RDT model, we will infer a solution W' to maximize the objective function $O(W)$ which is:

$$W' = argmax_W O(W)$$
(3)

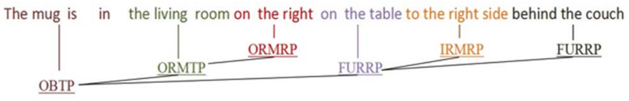
The expression l is used to get the shortest description and α is a constant number for the objective function. An RDT node y_k will selected to generate the spatial description if w_k=1.

D. Generate Natural Language

After inferring the best RDT model, the last step is the transition from the RDT node chain to the natural language description presenting the location of the target object based on the information in the RDT model. Considering the diversity and uncertainty of human-like spatial language, it is impossible to use a fixed prototype framework on language generation. Inspired by the previous work in [4], we consider the output natural language description as a tree structure being constructed by several clauses. An example of a tree-structured description is shown in Fig.3, which shows a language model grouping words into chunks (word phrases). Each chunk consists of a clause c and a chunk type name η. The type names and explanations are also shown in Fig.3. The potential relations that chunk A can have to chunk B include six possibilities: *neighbor-left(NL)*, *neighbor-right(NR)*, *parent-left(PL)*, *parent-right(PR)*, *child-left(CL)*, *child-right(CR)*. Assuming we have already inferred the best grounding model, which includes an RDT chain $y=\{y_1,\dots y_J\}$ including J $(J \le K)$ RDT nodes, let $Y=\{\upsilon_{\eta 1 \eta 2},\dots,\upsilon_{\eta J-1 \eta J}\}$ where $\upsilon_{\eta A \eta B} \in \{NL,NR,PL,PR,CL,CR\}$ is the relation between any two chunks A and B. Then the language generation work is to determine the set Y of all the chunks that can maximize the probability to generate a tree structure, which can be written as:

$$P(\Upsilon) = P(\{(c_1,\eta_1),y_1\},\dots\{(c_J,\eta_J),y_J\},\Upsilon) =$$
$$\prod_{j=1}^{J} P(c_j,\eta_j|y_j) \prod_{j1=1}^{J}\prod_{j2=1}^{J} P\left(\upsilon_{\eta_{j1}\eta_{j2}},|\eta_{j1}\eta_{j2},j1 \neq j2\right) \quad (4)$$

The conditional distribution can be easily trained by the corpus described in [2][4] which contains 149 template descriptions that were derived directly from older adult spatial language descriptions. The result will be determined not only based on the words and tag of each grounding unit but also on their sequence and structure. This enables the system to mimic a human-like style in spatial language descriptions.


The mug is in the living room on the right on the table to the right side behind the couch
OBTP
ORMTP
ORMRP
FURRP
IRMRP
FURRP

Outside Room Target Phrase	ORMTP
Outside Room Reference Phrase	ORMRP
Object Target Phrase	OBTP
Furniture Target Phrase	FURTP
Furniture Reference Phrase	FURRP
Inside Room Reference Phrase	IRMRP

Fig.3 A chunking tree structure of a spatial description and the explanation of the chunk names.

III. Evaluation Design

To examine our system, an experiment will be performed first in a simulated indoor environment which includes two rooms and a hallway between them (as shown in Fig.2). This setting has been used in our previous work on spatial language grounding and matches our physical lab space so that experiments can also be run in the real world environment. In a language generation test, the robot is initially placed in the hallway and then starts to search for a target object after it receives the object name from the human user. It will keep on roaming in the working environment and builds the environment model until it finds the target. The target object can be placed in one of six different places. For each path, a static and natural spatial language description will be generated. We will employ volunteer test subjects to score the spatial descriptions that are generated by robot.

IV. Conclusion

The development of this blueprint was an effort to achieve a natural spatial language generation system. Our project made apparent some of the challenges. The results of the early experiments confirm a decision to not use language templates but rather to use a human spatial language corpus to program a language generator.

Acknowledgment

This work was funded by the NSF under grant IIS-1017097.

References

[1] Scopelliti, M., M.V. Giuliani, and F. Fornara, Robots in a domestic setting: a psychological approach. Universal Access in the Information Society, 2005. 4(2): p. 146-155.

[2] Huo, Z., T. Alexenko, and M. Skubic. Using spatial language to drive a robot for an indoor environment fetch task. in Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on. 2014. IEEE.

[3] Skubic, M., Carlson, L., Miller, J., Li, X. O., and Huo, Z., Spatial language experiments for a robot fetch task. In Human-Robot Interaction (HRI), 2012 7th ACM/IEEE International Conference on (pp. 239-240). IEEE.

[4] Alexenko, T., Skubic, M., and Huo, Z. "Spatial Language Processing for Assistive Robots with" Deep" Chunking and Semantic Grammars." Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence. 2014.