Indoor Scene Knowledge Acquisition using a Natural Language Interface

Saranya Kesavan and Nicholas A. Giudice

School of Computing and Information Science
Spatial Informatics Program, University of Maine

nicholas.giudice@smaine.edu

Abstract. This paper proposes an interface that uses automatically-generated Natural Language (NL) descriptions to describe indoor scenes based on photos taken of that scene from smartphones or other portable camera-equipped mobile devices. The goal is to develop a non-visual interface based on spatio-linguistic descriptions which could assist blind people in knowing the contents of an indoor scene (e.g., room structure, furniture, landmarks, etc.) and supporting efficient navigation of this space based on these descriptions. In this paper, we concentrate on understanding the most salient content of a stereotypic indoor scene that is described by an observer, categorizing the description strategies employed in this process, and evaluating the best presentation of directional information using NL descriptions in order to support the most accurate spatial behaviors and mental representations of these scenes by means of human behavioral experiments. This knowledge will then be used to develop a domain specific indoor scene ontology, which in turn will be used to generate automated NL descriptions of indoor scenes based on their photographs, which will finally be integrated into a real-time non-visual scene description system.

Keywords: Natural Language, indoor scene description, indoor spatial knowledge, indoor scene ontology.

1 Introduction

Navigation involves a process of controlling and monitoring the movement of any physical entity from one place to another [1]. Humans carry out this navigation process in both outdoor and indoor environments, often with the aid of external navigation aids, such as maps or GPS-based guidance systems. While humans spend approximately 87% of their time indoors, comprising both familiar and unfamiliar indoor environments [2], real-time guidance systems only work outdoors due to attenuation of the GPS signal inside and a lack of standards for building information models [3]. Compared to outdoor travel, the lack of global landmarks and complexity of indoor environments makes the task of navigation within buildings more challenging, even with the advent of indoor navigation assistance [4]. The systems for navigation assistance that do exist are almost exclusively based on visual interfaces and thus are inaccessible to blind and low vision people, one of the fastest growing demographics of our aging population [5]. To address this information gap, this paper discusses the development of an indoor navigation and scene description system which provides

non-visual access to indoor environmental information by means of Natural Language descriptions delivered with the help of smartphones.

2 Background

The majority of the extant literature and technology development on accessible navigation devices relates to technology for detecting and avoiding obstacles to the path of travel or speech-enabled GPS systems for street navigation (see [6] for review). Several systems providing non-visual access to indoor environments have also been developed [see 7 for discussion]. As with the outdoor systems based on street networks, these technologies only provide network information about corridor connectivity or give landmark descriptions [8]. Beyond providing a label for a specific location (e.g., auditorium), there is no existing technology that describes the layout or salient features of these locations (i.e., the nodes which are linked by the network). As such, a blind person may have navigation assistance when traveling a route but once they reach their destination, access to functionally useful spatial knowledge regarding this location is generally limited or non-existent (e.g., the bounding contour of a meeting room, the position and orientation of a couch in the waiting room of a doctor's office, etc.).

Where navigation assistance of the network structure of large-scale environments benefits both blind and sighted users alike, sighted individuals rarely need assistance gaining information about these small-scale environments as the information is directly perceived through visual access to the scene. However, for a blind navigator relying on non-visual sensing, which is generally more proximal and less spatially precise, we argue that lack of access to spatial information about these local environments can be equally detrimental to accurate navigation, spatial learning, and cognitive map development. To date, limited research has been conducted to investigate the description of indoor scenes or how knowledge of the spatial distribution of architectural elements and salient objects in these spaces can be best imparted to blind people through non-visual channels. The limited research that has been done in this domain has required the use of expensive wearable specialty devices for acquiring spatial information. For example [9], discusses the use of a wearable device which converts visual information into tactile signals but carrying a specialized device for this purpose requires cumbersome, expensive hardware and the use of potentially confusing sensory translation algorithms. By contrast, our goal is to develop a system based on commercially-available hardware and an intuitive, easy to understand user interface. To this end, this paper proposes a work-in-progress system that provides non-visual access to specific indoor locations (scenes) through the use of NL descriptions delivered via a smartphone.

3 Natural Language – A Limited Information Display

Natural Language represents an intuitive interface as it is innate to most humans and is easy to generate using text-to-speech engines. It is often used in spatial contexts, e.g. direction giving, and has the advantage of being equally accessible to both sighted and blind users. Owing to the sparse information content that can be specified

using a serial, temporally extended, and low-bandwidth medium, NL is considered a limited information display. In addition, NL involves more cognitive load in working memory than perceptual interfaces as it requires cognitive mediation to interpret the verbal information being described, such as metric, topological, and other spatial information [10]. This paper proposes a way to effectively use this limited information medium in an accessible scene description system for blind users.

4 Behavioral Experiments

Because of recent technological advancements and promising results in the field of Natural Language processing and generation, we argue that NL is one of the most important modes of information access for incorporation in non-visual interfaces and intelligent devices used by blind people. One problem is that NL description generation primarily concentrates on the semantic and syntactic aspects of the linguistic description in order to mimic human speech patterns. However, there is little formal research on understanding the ways in which a human summarizes the information they directly perceive, especially when looking at an indoor scene, through spatial verbal descriptions in order to convey survey knowledge of the scene.

Hence, in order to generate a NL description of an indoor scene, we argue that it is important to first understand the ways in which humans would naturally describe (e.g., verbally narrate) the space. NL generated without understanding of this narration is only a formal arrangement of words into sentences which abides the syntactic and semantic rules of a language following a specific architecture. To gain this knowledge, behavioral experiments must be conducted in order to understand the logic behind the human-generated NL description of an indoor scene. These results can then be compared to descriptions generated by a machine-generated NL description to assess where differences and similarities arise. The following human experiments are proposed to address this question.

Direct Observation versus Photographic Observation of an indoor scene

The end goal is for blind persons to use photos taken with their smartphones in order to obtain information about the spatial configuration of indoor scenes, including the location of its constituent objects, delivered via NL descriptions. Photos taken using smartphone cameras will inherently have a limited field of view (FOV). Hence it is important to compare the spatial information obtained from photographic observations of an indoor scene against the spatial information obtained from direct observations of the same scene to evaluate whether this limited FOV leads to exclusion of important environmental details in the ensuing spatial verbal descriptions. A behavioral experiment was conducted to evaluate whether there is a significant difference of observation by comparing the accuracy of scene re-creation based on previously generated scene descriptions from both modes. Supporting the efficacy of camera-based photos in our system, results revealed no significant differences between spatial information acquired from human or camera-based observations or re-creation accuracy based on descriptions generated from these two modes [11].

Comparing Description Strategies

Flexibility is one of the most important features of a Natural Language. One challenge is that NL descriptions of spatial information of objects in an indoor scene could be structured in different ways following different strategies. For example, a description could begin by describing the name and spatial locations of objects in one corner of the room and then follow a cyclic clockwise strategy of describing the other objects around the room. Alternatively, a description could combine objects based on their functionality, e.g. describing the spatial location of all the tables that are present in a room, then describing the chairs, etc.

Research conducted in [12] suggests that the choice of description strategies also depends on the spatial extent being described. Hence, it is important to first identify the different scene description strategies people adopt from a common perspective and then to determine which of these strategies leads to the acquisition of the most accurate spatial information while minimizing the cognitive effort required for the process. A behavioral study is currently being conducted to understand the different types of strategies that are used by humans, and the effectiveness of each for conveying accurate indoor scene descriptions. Another behavioral study will then be conducted to investigate which among those strategies helps the user to gain the most accurate spatial information supporting spatial learning and behavior in the space.

Presentation of Directional Cues

For a linguistic scene description to generate a mental map that is comparable in function with mental maps developed from visual perception, it is extremely important to have an accurate method for specifying directional cues about the spatial locations of objects within the scene. As with scene description strategies, there are different ways to verbally present these directional cues to the user. The work done in [13] suggests that people best understand directional cues when they are presented using relative directions rather than using only absolute directions. However, we are unaware of formal research investigating the best way to present directional cues with the highest precision within a relative reference frame.

Degree measurements and clock face directions are the most common ways to present angular information using a relative frame of reference. For example, "a desk is at your 1 O' clock position" and "a desk is at 30 degrees on your right" both specify the same spatial location of the desk. But it is important to know which of these presentation methods leads to the most accurate perception of directional information. To address this question, a behavioral study is currently being conducted to compare the accuracy of angular perception based on these two types of directional cues.

5 Machine Generated vs. Human Generated Natural Language

Natural Language descriptions of indoor scenes will be generated based on the results of the above mentioned behavioral experiments. They are expected to provide access to accurate spatial information for use by a blind user when made available. However, it is also important to compare the NL descriptions created by an automated machine with the NL descriptions created by a human user in order that results from the latter can guide development of the former. This could be tested by asking partici-

pants to reproduce the scenes based on the NL descriptions from both human generated and machine generated descriptions. The accuracy of re-created scenes should be tested for any significant differences in the ensuing re-creations before being implemented in a real time indoor scene description system.

6 Indoor Scene Ontology

Natural Language Generation architecture involves a procedural and formal way of arranging raw spatial information that must then be converted to a NL [14]. To support this process, it is important to represent spatial information of indoor scenes in a formal setting, e.g. as an indoor scene ontology. Although we have ontologies available for characterizing indoor spaces in terms of corridors and pathways [15], there are currently no ontologies available to represent indoor scenes. We argue for the importance of constructing an indoor scene ontology which represents human described scene information. The primary goal of this ontology is to formally reflect and represent the ways in which humans perceive space (from the above mentioned behavioral experiments) and to structure the relevant information into a robust and flexible NL description. For example, the envisaged indoor scene ontology should involve a saliency rating of the objects that are typically present in an indoor scene. It should also be related with the existing linguistic ontology of space as proposed in [16] in order to fill the gap between human perception and formal linguistic procedures used in NL generation. Using the NL Generation techniques mentioned in [14], a NL description of indoor scenes could be developed based on the information represented in the proposed indoor scene ontology.

7 Conclusion

This paper proposes a NL user interface for describing indoor scenes to visually impaired people. While current natural language systems concentrate on the semantic and syntactic components of natural language, we propose an automated NL system that is aimed at mimicking accurate descriptions delivered by humans in terms of spatial knowledge acquisition and information delivery as established from our human behavioral experiments. Several experiments are proposed to better understand the ways in which humans perceive and describe indoor scenes in order to establish the most salient information content and description strategies. Finally, we propose the construction of an indoor scene ontology to formally represent the knowledge acquired from the results of our behavioral experiments.

8 Acknowledgement

This project was supported by NSF grant CDI-1028895. Thanks also to Bill Whalen for comments on the manuscript.

References

1. Bowditch, N.: CHAPTER 1 INTRODUCTION TO MARINE NAVIGATION. The American Practical Navigator. pp. 1-14. Defense Mapping Agency Hydrographic/Topographic Center, Maryland (1995).
2. Klepeis, N.E., Nelson, W.C., Ott, et al.: The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. Journal of Exposure Analysis and Environmental Epidemiology. 11, 231-52 (2001).
3. Giudice, N.A, Walton, L & Worboys, M.F.: The informatics of indoor and outdoor space: A research agenda. Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness. 47-53 (2010).
4. Butz, A., Kr, A., & Lohse, M.: A Hybrid Indoor Navigation System. Proceedings of the 6th international conference on Intelligent user interfaces (IUI '01). 25-32 (2001).
5. World Health Organization: WHO Visual impairment and blindness, http://www.who.int/mediacentre/factsheets/fs282/en/.
6. Giudice, N.A., & Legge, G.E.: Blind navigation and the role of technology. Handbook of Smart Technology for Aging, Disability, and Independence. (2008).
7. Kulyukin, V., Gharpure, C., Nicholson, J., Pavithran, S.: RFID in robot-assisted indoor navigation for the visually impaired. 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2, 1979-1984 (2004).
8. Bentzen, B. & Mitchell, P. A.: Audible Signage as a Wayfinding Aid: Verbal Land mark versus Talking Signs. Journal of Visual Impairment & Blindness, 89, 6, 494-505 (1995).
9. Johnson, L. A, & Higgins, C.M.: A navigation aid for the blind using tactile-visual sensory substitution, Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 1, 6289-92 (2006).
10. Klatzky, R.L., Marston, J.R., Giudice, N. A, Golledge, R.G., Loomis, J.M.: Cognitive load of navigating without vision when guided by virtual sound versus spatial language. Journal of experimental psychology, Applied. 12, 223-32 (2006).
11. Kesavan, S., & Giudice, N.A.: Automated natural language description of indoor spaces. Proceedings of the doctoral Colloquium, COSIT '11. 1-5 (2011).
12. Tenbrink, T., & Coventry, K.: Spatial Strategies in the Description of Complex Con figurations. Discourse Processes: A multi-disciplinary journal. 8, 37-41 (2011).
13. Ishikawa, T., & Kiyomoto, M.: Turn to the Left or to the West : Verbal Navigational, Proceedings of the 5th GIScience international conference, 5266, 119-132 (2008).
14. Reiter, E. & Dale, R.: Building applied natural language generation systems. Natural Language Engineering, 3, 1, 57-87 (1997).
15. Worboys, M.: Modeling indoor space. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Indoor Spatial Awareness - ISA '11. 1-6 (2011).
16. Bateman, J. A., Hois, J., Ross, R., & Tenbrink, T.: A linguistic ontology of space for natural language processing. Artificial Intelligence. 174, 1027-1071 (2010).