Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 48 (2018) 1–14

www.elsevier.com/locate/csl

Situated reference resolution using visual saliency and crowdsourcing-based priors for a spoken dialog system within vehicles☆

Teruhisa Misu

Honda Research Institute USA, 375 Ravendale Drive Mountain View, CA 94043 USA

Received 18 January 2016; received in revised form 5 April 2017; accepted 28 September 2017
Available online 12 October 2017

Abstract

In this paper, we address issues in situated language understanding in a moving car. More specifically, we propose a reference resolution method to identify user queries about specific target objects in their surroundings. We investigate methods of predicting which target object is likely to be queried given a visual scene and what kind of linguistic cues users naturally provide to describe a given target object in a situated environment. We propose methods to incorporate the visual saliency of the visual scene as a prior. Crowdsourced statistics of how people describe an object are also used as a prior. We have collected situated utterances from drivers using our research system, which was embedded in a real vehicle. We demonstrate that the proposed algorithms improve target identification rate by 15.1% absolute over the baseline method that does not use visual saliency-based prior and depends on public database with a limited number of category information.

Keywords: Situated dialog; In-car interaction; Visual saliency; Crowdsourcing; Multimodal interaction

1. Introduction

Recent advances in sensing technologies have allowed researchers to explore applications that require clear awareness of a system's dynamic context and physical surroundings. Such applications include multi-participant conversation systems (Bohus and Horvitz, 2009) and human-robot interaction systems (Tellex et al., 2011; Sugiura et al., 2011). The general problem of understanding and interacting with human users in such environments is referred to as *situated interaction*. In this study, we address an environment in which situated interactions take place, i.e., a moving car. Previous studies that have analyzed in-car interactions between an expert co-pilot and a driver (Cohen et al., 2014) have shown that people frequently use referring expressions about their surroundings (e.g., *What is that big building on the right?*). In addition, conversations between a driver and a passenger focus on the surrounding objects. (e.g., *Take a left turn in front of that blue restaurant.*). Therefore, we believe that a method to find a link between the user utterance and physical object in the surrounding

☆ This paper has been recommended for acceptance by Prof. R. K. Moore.

E-mail address: teruhisa.misu@gmail.com

http://dx.doi.org/10.1016/j.csl.2017.09.001

environment (situated reference resolution) is a key technology to facilitate driver-vehicle and driver-passenger communications.

Data fusion from various media inputs is vital to achieve such reference resolution for in-car interactions. Cohen et al. (2014) showed that many people do not necessarily present a detailed description of the target, but simply say "what's that". Drivers use multimodal cues such as head-pose or gestures to indicate their target. In practice, however, head-pose and/or eye-gaze signals can be problematic. First, head-pose or eye-gaze detection in changing lighting conditions (e.g., outdoors) is difficult, because the estimation algorithm relies on a camera image. Second, mapping between eye-gaze behavior and user intention is not straightforward in interactions in a rapidly changing environment while driving, in which eye fixations are unusual activities. Third, eye-gaze estimation fundamentally requires calibration because the optical axis of the human eye differs by individual offset from the visual axis, which is the true gaze direction (Hansen and Ji, 2010). Therefore, a calibration-free method is preferred.

Linguistic information, such as the color and position of the target object is one of the most informative cues for situated reference resolution; however, such linguistic information often lacks practicality when applied to real-world situation. In such situations, problem setting is not necessarily limited to pre-defined situations, color information is often difficult due to changes of lighting conditions. Computer vision-based techniques have been applied to acquire property/category information of objects in an environmental context. Most conventional studies have considered relatively simple scenes containing single color objects without occlusions or complex backgrounds. However, scene recognition in the real world remains challenging and error-prone. Although public point-of-interest (POI) databases (DB) (e.g., The Google Places API[1] and Yelp API[2]) include properties for search purposes (i.e., name, geolocation, category, and cuisine), the DBs do not include properties that people are likely to use in an environmental context (e.g., color, equipment, and position). In addition, subjective information (e.g., size) is difficult to obtain with image recognition techniques.

We have also implemented in-car spoken dialog system with a reference resolution function (Misu et al., 2014; Kim and Misu, 2014; Misu et al., 2015). In those studies, we have investigated the effectiveness of driver head-pose information. However, sensor calibration and adaptation to drivers (and driving position) were required to get effective improvement for a reference resolution task in a moving car (Kim and Misu, 2014). We also confirmed that linguistic information (Misu et al., 2014) assuming that all properties concerning the target objects are available. However, preparing such database is costly.

To address the above issues, in the present work, we introduce two priors for in-car situated dialog systems. One is based on visual saliency, which is defined as a quality that makes some region of an image stand out relative to its neighbors. We assume that one reason we sometimes do not necessarily use explicit linguistic/multimodal cues is that we can achieve saliency-based joint attention naturally. We incorporate visual saliency as a directional prior. The other prior used in this study is linguistic cues that people are likely to provide given the POI in an environmental context. We use crowdsourcing to collect such statistics.

In summary, we address the following research questions in dealing with situated reference resolution in vehicles, and demonstrate the effectiveness of the proposed methods through data collection and analyses.

1. Can we determine a user calibration-free multi-modal cue to identify driver intention using visual saliency?
2. Is crowdsourcing effective to obtain the categories/properties of real-world objects? How useful is the data in a target identification task compared to public and ideal DBs?

The paper is organized as follows. We cover recent related work in Section 2. Followed by the overview of the Townsurfer in-car spoken dialog system, we propose methods to improve situated language understanding in Section 3. Based on the collected data described in Sections 4, we analyze the contributions of the proposed methods in Section 5. We then clarify our research contributions through discussion in Section 6. Finally, we conclude our work in Section 7.

[1] https://developers.google.com/places/.
[2] http://www.yelp.com/developers/.