Using a Spoken Dialogue System for Crowdsourcing Street-level Geographic Information

Raveesh Meena, Johan Boye, Gabriel Skantze, Joakim Gustafson
KTH Royal Institute of Technology
School of Computer Science and Communication, Lindstedtsvägen 3, 10044 Stockholm, Sweden
{raveesh, jboye}@csc.kth.se, {gabriel, jocke}@speech.kth.se

Abstract

We present a novel scheme for enriching geographic database with street-level geographic information that could be useful for pedestrian navigation. A spoken dialogue system for crowdsourcing street-level geographic details was developed and tested in an in-lab experimentation. The system obtained 96.4% of concept values correctly after interacting with the first six of the fifteen users. This indicates that the proposed scheme holds promise for a wider real life application.

1. Motivation

Studies have shown that inclusion of *landmarks* (distinctive objects in the city environment) into routing instructions for pedestrians raises the user's confidence in the system, compared to only left-right instructions typically used for car navigation (Ross et al., 2004). The pedestrian navigation system presented in Boye et al. (2014) follows this strategy and could generate instructions like "go towards the *SEB bank*," for the routing scenario in Figure 1. However, when comparing the map with the street-view picture in Figure 2, it becomes obvious that the "SEB" bank office is very hard to see and probably not very suitable to use as a landmark in route descriptions. On the other hand, an instruction like "go towards the *yellow* building" could have been easily resolved by the user. However, the geographic database (OpenStreetMap in case of this system) does not contain the fact that the building has six stories and a façade made of yellow bricks. This is not due to any shortcoming of the database; it just goes to show that the database has been constructed with map drawing in mind, rather than pedestrian routing.

Landmark based pedestrian routing requires that the geographic database contains many landmarks and many details about them. Current crowdsourcing schemes for mapping, however, are inadequate for adding street-level details, such as the presence of signs on the façade of buildings. These details are hard to add offline, sitting in front of one's PC using a map interface. We think using a spoken dialogue system for crowdsourcing such details – while the user is out there on the streets – is an interesting application of speech technology. Not only is using speech convenient, the users could be asked to freely describe objects they consider important in their current view. In this way, the system could learn new objects not anticipated by the system designers, and their associated properties.

To investigate the potential of these ideas we developed a spoken dialogue system for crowdsourcing street-level geographic information and tested it in in-lab experimentation (Meena et al., 2014). The approach has shown that it is possible to leverage a spoken dialogue system to obtain street-level details from users. In this paper we present a brief overview of this proof-of-concept study and present only one of the various analyses of the results from the main paper.

Figure 1: A pedestrian routing scenario (the blue arrow indicates the position and direction of the user)

Figure 2: The visual scene corresponding to the pedestrian routing scenario in Figure 1

2. A dialogue system for crowd-sourcing

We wanted to investigate if a dialogue system can be used to obtain details about landmarks in a map (such as Figure 1) from users who have access to the corresponding visual scene (street-level first-person picture, such as Figure 2). Following the observations of a role-playing experiment in which one person acted as the system and asked the other person (the user) questions to seek details about landmarks in a map, we designed our dialogue system to operate in two modes of interaction: slot-filling and open-ended. In the slot-filling mode the system takes and retains initiative during the interaction and asks *wh*–questions, acknowledges or clarifies (*yes-no* questions) landmark specific properties (e.g. colour, number of floors, and presence of signs). During the open-ended interactions, the system behaves as active

listener and asks questions like "Could you describe the sign on the façade?" or "Have I missed any important landmarks in this scene?"

The system is fully autonomous: for a given visual scene it first obtains basic details such as geographic co-ordinates of the landmarks and their type. Towards this it uses the Visibility Engine, presented in Boye et al. (2014), which scans the OpenStreetMap database for a specified viewport and depth. The dialogue manager then uses these basic details to generate referential expressions, such as "Do you see a *building* on *your left*?" and attempts to ground the target landmark with the user. Following a successful grounding the system iterates over a list of predefined properties (slots) for that landmark type and engages in spoken dialogue to learn the values.

We also wanted the system to be able to learn the most reliable hypothesis for slot values. To achieve this, the system uses an XML encoding for visual scenes in which it stores the mapped landmarks, their properties, and all the various slot values obtained from different users. At the end of each interaction the system updates its beliefs about slot values by assuming the slot values with majority vote as the most reliable estimate. The system could then engage with new users in a clarification type dialogue to verify whether the learned estimates are valid.

3. In-lab crowd-sourcing experiment

Nine visual scenes (taken in Stockholm city, cf. Figure 2) were used for the task of crowd-sourcing. Fifteen human subjects (4 females and 11 males) took part in the crowdsourcing exercise. Participants were placed in front of a computer display and were told that the system will engage them in a spoken conversation to seek or clarify details about landmarks and other objects in the visual scenes. They were told that the details would be used for pedestrian routing and therefore they are free to choose the details they thought could be useful for navigation.

For this experiment, the system engaged users in either wh– or yes–no questions only for a visual scene. The number of wh– and yes–no interactions were balanced. However, visual scenes presented with wh– queries to a participant p_i were presented to participant p_{i+1} with only yes–no queries (clarifications), and vice-versa. Thus while the system uses wh– type interactions to gain information and update system beliefs, the yes–no interactions were used to verify (and update) the beliefs with the next user. Table 1 presents an example wh– type interaction (except for the opening yes-no statement for *grounding*) for the visual scene in Figure 2.

Sys	yn– query	Do you see a building on the far left?
Usr		Yes
Sys	Backchannel	Mh-hmm
Sys	wh– query	What color does the building have?
Usr		Cream color [Green color]
Sys	Reprise	Green
Sys	wh– query	How many floors is this building?
Usr		I don't know [I do no]
	…	
Sys	wh– query	Anything peculiar in this scene?
Usr		There is a bike painted on street just in front of me [is the right pink university in front of may]
Sys	Backchannel	Mh-hmm
Usr		That's it [that that]
Sys	Ending	Okay, thanks

Table 1: Sample interaction with wh– queries only

4. Data analysis

A total of 197 slots were learned in the experiment. We analysed how many slot-values had been correctly retrieved after 1, 2… 15 users. In Figure 3, the curve "Majority" illustrates the fraction of slot-values correctly learned with each new user, under the assumption that the slot-values with majority votes – from all the 15 users – constitute the ground truth. Thus after interacting with the first user the system had obtained 67.0% of slot-values correctly and 20.8% incorrectly (Not-in-Majority), whereas 12.2% slots were not queried suggesting that some landmarks were not grounded by the users). Interactions with the first six users only were sufficient to obtain 96.4% of slot-values correctly.

We also investigated how close the majority is to the actual truth. One of the co-authors labelled all the obtained slot-values as *sensible* or *insensible*, based on the combined knowledge from the corresponding maps, the visual scenes, and the set of obtained values. The progression curves "Sensible" and "Insensible" in Figure 3 illustrate the fraction of total slots for which the learned values were *actually* correct and incorrect, respectively. Interactions with only first six users were able to obtain 80.0% of slot-values with actually true values. These results are promising and encourage a real life experimentation with real users out on the streets.

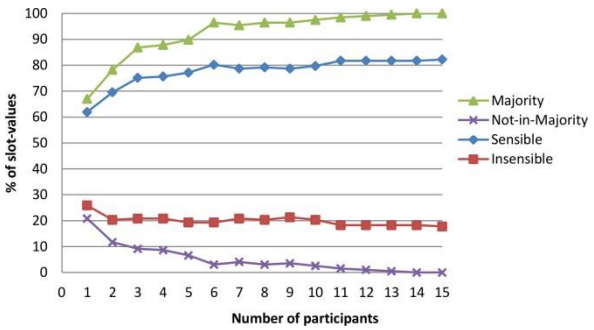

Figure 3: Rate of learning slot-values

Reference

Boye, J., Fredriksson, M., Götze, J., Gustafson, J., & Königsmann, J. (2014). Walk This Way: Spatial Grounding for City Exploration. In Mariani, J., Rosset, S., Garnier-Rizet, M., & Devillers, L. (Eds.), *Natural Interaction with Robots, Knowbots and Smartphones* (pp. 59-67). Springer New York.

Meena, R., Boye, J., Skantze, G., & Gustafson, J. (2014). Crowdsourcing Street-level Geographic Information Using a Spoken Dialogue System. In *15th Annual Meeting of the Special Interest Group on Discourse and Dialogue - SIGdial*. Philadelphia, PA, US.

Ross, T., May, A., & Thompson, S. (2004). The Use of Landmarks in Pedestrian Navigation Instructions and the Effects of Context. In Brewster, S., & Dunlop, M. (Eds.), *Mobile Human-Computer Interaction - MobileHCI 2004* (pp. 300-304). Springer.