Crowdsourcing Street-level Geographic Information Using a Spoken Dialogue System

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

Abstract

We present a technique for crowdsourcing street-level geographic information using spoken natural language. In particular, we are interested in obtaining first-person-view information about what can be seen from different positions in the city. This information can then for example be used for pedestrian routing services. The approach has been tested in the lab using a fully implemented spoken dialogue system, and has shown promising results.

1 Introduction

Crowdsourcing is increasingly being used in speech processing for tasks such as speech data acquisition, transcription/labeling, and assessment of speech technology, e.g. spoken dialogue systems (Parent & Eskenazi, 2011). However, we are not aware of any attempts where a dialogue system is the *vehicle* for crowdsourcing rather than the object of study, that is, where a spoken dialogue system is used to collect information from a large body of users. A task where such crowdsourcing dialogue systems would be useful is to populate geographic databases. While there are now open databases with geographic information, such as OpenStreetMap (Haklay & Weber, 2008), these are typically intended for map drawing, and therefore lack detailed street-level information about city landmarks, such as colors and height of buildings, ornamentations, facade materials, balconies, conspicuous signs, etc. Such information could for example be very useful for pedestrian navigation (Tom & Denis, 2003; Ross et al., 2004). With the current growing usage of smartphones, we might envisage a community of users using their phones to contribute information to geographic databases, annotating cities to a great level of detail, using multi-modal method including speech. The key reason for using speech for map annotation is convenience; it is easy to talk into a mobile phone while walking down the street, so a user with a little experience will not be slowed down by the activity of interacting with a database. This way, useful information could be obtained that is really hard to add offline, sitting in front of one's PC using a map interface, things like: Can you see X from this point? Is there a big sign over the entrance of the restaurant? What color is the building on your right?

Another advantage of using a spoken dialogue system is that 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.

In this paper we present a proof-of-concept study of how a spoken dialogue system could be used to enrich geographic databases by crowdsourcing. To our knowledge, this is the first attempt at using spoken dialogue systems for crowdsourcing in this way. In Section 2, we elaborate on the need of spoken dialogue systems for crowdsourcing geographic information. In Section 3 we describe the dialogue system implementation. Section 4 presents our in-lab crowdsourcing experiment. We present an analysis of crowd-sourced data in Section 5, and discuss directions for future work in Section 6.

2 The pedestrian routing domain

Routing systems have been around quite some time for car navigation, but systems for pedestri-

Proceedings of the SIGDIAL 2014 Conference, pages 2–11,
Philadelphia, U.S.A., 18-20 June 2014.

an routing are relatively new and are still in their nascent stage (Bartie & Mackaness, 2006; Krug et al., 2003; Janarthanam et al., 2012; Boye et al., 2014). In the case of pedestrian navigation, it is preferable for way-finding systems to base their instructions on *landmarks*, by which we understand distinctive objects in the city environment. Studies have shown that the inclusion of landmarks into system-generated instructions for a pedestrian raises the user's confidence in the system, compared to only left-right instructions (Tom & Denis, 2003; Ross et al., 2004).

Basing routing instructions on landmarks means that the routing system would, for example, generate an instruction "Go towards the red brick building" (where, in this case, "the red brick building" is the landmark), rather than "Turn slightly left here" or "Go north 200 meters". This strategy for providing instructions places certain requirements on the geographic database: It has to include many landmarks and many details about them as well, so that the system can generate clear and un-ambiguous instructions. However, the information contained in current databases is still both sparse and coarse-grained in many cases.

Our starting point is a pedestrian routing system we designed and implemented, using the landmark-based approach to instruction-giving (Boye et al., 2014). The system performs visibility calculations whenever the pedestrian approaches a waypoint, in order to compute the set of landmarks that are visible for the user from his current position. OpenStreetMap (Haklay & Weber, 2008) is used as the data source. Figure 1 shows a typical situation in pedestrian routing session. The blue dot indicates the user's position and the blue arrow her direction. Figure 2 shows the same situation in a first-person perspective. The system can now compute the set of visible landmarks, such as buildings and traffic lights, along with distances and angles to those landmarks. The angle to a building is given as an interval in degrees relative to the direction of the user (e.g. 90° left to 30° left). This is exemplified in Figure 1, where four different buildings are in view (with field of view marked with numbers 1–4). Landmarks that are not buildings are considered to be a single point, and hence the relative angle can be given as a single number.

When comparing the map with the street view picture, 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, the database does not contain the fact that the building has six stories and a façade made of yellow bricks, something that would be easily recognizable for the pedestrian. This is not due to any shortcoming of the OpenStreetMap database; it just goes to show that the database has been constructed with map drawing in mind, rather than pedestrian routing. There are also some other notable omissions in the database; e.g. the shop on the corner, visible right in front of the user, is not present in the database. Since OpenStreetMap is crowd-sourced, there is no guarantee as to which information will be present in the database, and which will not. This also highlights the limitation of existing approaches to crowd-sourcing geographic information: Some useful information is difficult to add off-line, using a map interface on a PC. On the other hand, it would be a straightforward matter given the kind of crowd-sourcing spoken dialogue system we present next.

Figure 1: A pedestrian routing scenario

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

3 A dialogue system for crowd-sourcing

To verify the potential of the ideas discussed above, we implemented a spoken dialogue system that can engage in spoken conversation with

users and learn details about landmarks in visual scenes (such as Figure 2). To identify the kind of details in a visual scene that the system could potentially ask the users, we first conducted a preliminary informal crowd-sourcing dialogue: one person (the receiver), was instructed to seek information that could be useful for pedestrian navigation from the other person (the giver). The receiver only had access to information available in the maps from OpenStreetMap, as in Figure 1, but without any marking of field of views, whereas the giver only had access to the corresponding visual scene (as in Figure 2). Interaction data from eight such dialogues (from four participants, and four different visual scenes) suggested that in a city environment, buildings are prominent landmarks and much of the interaction involves their properties such as color, number of stories, color of roof, signs or ornamentations on buildings, whether it has shops, etc. Seeking further details on mentioned signs, shops, and entities (whether mapped or unmapped) proved to be a useful strategy to obtain information. We also noted that asking for open-ended questions, such as "*Is there anything else in this scene that I should be aware of*?" towards the end has the potential of revealing unknown landmarks and details in the map.

Obtaining specific details about known objects from the user corresponds to slot-filling in a dialogue system, where the dialogue system seeks a value for a certain slot (= attribute). By engaging in an open-ended interaction the system could also obtain general details to identify new slot-value pairs. Although slots could be in some cases be multi-valued (e.g., a building could have both color red and yellow), we have here made the simplifying assumption that they are single valued. Since users may not always be able to specify values for slots we treat *no-value* as a valid slot-value for all type of slots.

We also wanted the system to automatically learn the most reliable values for the slots, over several interactions. As the system interacts with new users, it is likely that the system will obtain a range of values for certain slots. The variability of the answers could appear for various reasons: users may have differences in perception about slot-values such as colors, some users might misunderstand what building is being talked about, and errors in speech recognition might result in the wrong slot values. Some of these values may therefore be in agreement with those given by other users, while some may differ slightly or be in complete contradiction. Thus the system should be able to keep a record of all the various slot-values obtained (including the disputed ones), identify slot-values that need to be clarified, and engage in a dialogue with users for clarification.

In view of these requirements, we have designed our crowd-sourcing dialogue system to be able to (1) take and retain initiative during the interactions for slot-filling, (2) behave as a responsive listener when engaging in open-ended dialogue, and (3) ask *wh*– and *yes–no questions* for seeking and clarifying slot-values, respectively. Thus when performing the slot-filling task, the system mainly asks questions, acknowledges, or clarifies the concepts learned for the slot-values. Apart from requesting repetitions, the user cannot ask any questions or by other means take the initiative. A summary of all the attributes and corresponding system prompts is presented in Appendix A.

The top half of Figure 3 illustrates the key components of the dialogue system. The Dialogue Manager queries the Scene Manager (SM) for slots to be filled or slot-values to be clarified, engages in dialogue with users to learn/clarify slot-values, and informs the SM about the values obtained for these slots. The SM manages a list of scenes and the predefined slots – for each type of landmark in visual scenes – that need to be filled, maintains a record of slot-values obtained from all the users, and identifies slot-values with majority vote as the current reliable slot-value. To achieve these objectives, the scene manager uses an XML representation of visual scenes. In this representation, landmarks (e.g., buildings, junctions, etc.) – automatically acquired through the OpenStreetMap database and the visibility computations mentioned in Section 2 – are stored as *scene-objects* (cf. Figure 4).

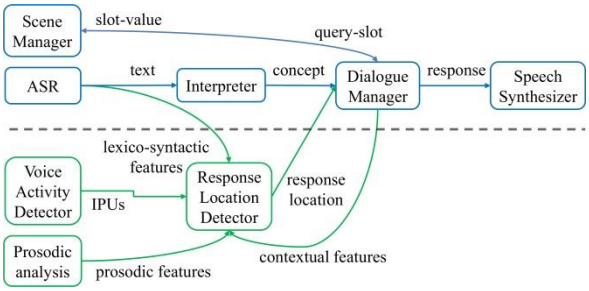

Figure 3: Dialogue system architecture

The Dialogue Manager (DM) uses scene-object attributes, such as type, angle or interval of a building, to generate referential expressions, such as "Do you see a *building* on the *far left*?"

or "Do you see a *shop* on the *left*?" to draw the users' attention to the intended landmark in the scene. During the course of interaction, the Scene Manager (SM) extends scene-objects with a set of predefined attributes (= slots) that we identified in the preliminary study, along with their various slot-values (cf. Figure 5). For each slot, the SM keeps a record of slot-values obtained through wh– questions as well as the ones disputed by the users in yes–no questions (cf. `obtained` and `disputed` tags in the XML), and uses their tally to identify the slot-value in majority. The system assumes this slot-value (or one of them in case of a tie) as its best estimate of a slot-value pair, which it could clarify with another user using a yes–no query. During the slot-filling mode the DM switches to open-ended interaction mode to seek general details (using prompts such as "*Could you describe it/them?*"), if the user suggests/agrees that there are signs on/at a scene-object, or a building has shops or restaurants. Once all the slots for all the scene-objects in a visual scene have been queried, the DM once again switches to the open-ended interaction mode and queries the users whether there are any other relevant signs or landmarks that the system may have missed and should be aware of. On completion of the open-ended queries the SM selects the next visual scene, and the DM engages in a new dialogue.

```
<scene xmlns="cityCS.scene" name=" view7.jpg" lat="59.34501"
lon="18.0614" fovl="-60" fovr="60" bearing="320" dist="100">
    <scene-object>
        <id>35274588</id> <type>building</type>
        <from>-60</from> <end>-39</end>
    </scene-object>
    <scene-object>
        <id>538907080</id> <type>shop</type>
        <distance>34.82</distance>
        <angle>-39</angle> <bearing>281</bearing>
    </scene-object>
    <scene-object>
        <id>280604</id> <type>building</type>
        <from>-38</from> <end>6</end>
    </scene-object>
    <scene-object>
        <id>193906</id> <type>traffic_signals</type>
        <distance>40.77</distance>
        <angle>-14</angle> <bearing>306</bearing>
    </scene-object>
    ...
</scene>
```

Figure 4: XML representation of visual scenes

For speech recognition and semantic interpretation the system uses a context-free grammar with semantic tags (SRGS[1]), tailored for the domain. The output of semantic interpretation is a concept. If the concept type matches the type of the slot, the dialogue manager informs the scene manager about the obtained slot-value. If the concept type is inappropriate the DM queries the user once more (albeit using different utterance forms). If still no appropriate concept is learned the DM requests the SM for the next slot and proceeds with the dialogue. For speech synthesis, we use the CereVoice system developed by CereProc[2]. The dialogue system has been implemented using the IrisTK framework (Skantze & Al Moubayed, 2012).

```
<scene-object>
    <id>35274588</id> <type>building</type>
    <from>-60</from> <end>-39</end>
    <slot slotName="VISIBLE">…    </slot>
    <slot slotName="COLOR">
     <obtained>
       <value slotValue="Green">
         <userlist>
           <usrDtls uid="u01" asrCnf="0.06" qType="WH"/>
         </userlist>
       </value>
       <value slotValue="no-value">
         <userlist>
           <usrDtls uid="u02" asrCnf="0.46" qType ="WH"/>
         </userlist>
       </value>
       <value slotValue="Gray">
         <userlist>
           <usrDtls uid="u03" asrCnf="0.19" qType ="WH"/>
         </userlist>
       </value>
     </obtained>
     <disputed>
       <value slotValue="Green">
         <userlist>
           <usrDtls uid="u02" asrCnf="0.92" qType ="YN"/>
         </userlist>
       </value>
     </disputed>
    </slot>
    <slot slotName="STORIES">…    </slot>
    <slot slotName="ROOF_COLOR">…    </slot>
    …
</scene-object>
```

Figure 5: Every slot-value is recorded

In contrast to the slot-filling mode, when engaging in an open-ended interaction, the system leaves the initiative to the user and behaves as a responsive listener. That is, the system only produces feedback responses, such as backchannels (e.g., *okay*, *mh-hmm*, *uh-huh*), repetition requests for longer speaker turns (e.g., *could you repeat that?*), or continuation prompts such as "*anything else?*" until the user is finished speaking. Unless the system recognized an explicit closing statement from the user (e.g., "I can't"), the system encourages the user to continue the descriptions for 2 to 4 turns (chosen randomly).

To detect appropriate locations in users' speech where the system should give feedback response, the system uses a trained data-driven model (Meena et al., 2013). When the voice activity detector detects a silence of 200 ms in users' speech, the model uses prosodic, contextual and lexico-syntactic features from the preceding speech segment to decide whether the system

[1] http://www.w3.org/TR/speech-grammar/

[2] https://www.cereproc.com/

should produce a feedback response. The lower half of Figure 3 shows the additional components of the dialogue system used in open-ended interaction mode. In this mode, the ASR system uses a language model that is trained on interactions from a related domain (verbal route descriptions), in parallel to the SRGS grammar.

4 In-lab crowd-sourcing experiment

Nine visual scenes (wide-angle pictures in first-person perspective and taken in Stockholm city, cf. Figure 2) were used for the task of crowdsourcing. Fifteen human participants (4 females and 11 males) participated in the crowdsourcing exercise. All participants either studied or worked at the School of Computer Science and Communication, KTH, Stockholm.

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 visual scenes. They were told that the details would be used for pedestrian routing and therefore they are free to choose and specify details (in open-ended questions) that they thought would be useful when giving route instructions to another person.

Each participant did the nine visual scenes in the same order, with a 1 minute pause between each of them. The first visual scene was used as a trial in order to familiarize participants with the interaction scenario. For this reason, the trial interaction was specifically designed to engage the participants in both wh– and yes–no type questions. We also wanted to investigate whether the use of wh– or yes–no questions alone has any impact on the interactions. For the remaining eight interactions, the dialogue system therefore alternatively used wh– and yes–no questions throughout the whole dialogue. The order of query types was switched for each successive participant. Thus visual scenes presented with wh–queries to a participant p_i were presented to participant p_{i+1} with only yes–no queries, and vice-versa. After each interaction participants were asked (1) whether the system appeared to be familiar with the visual scene; and (2) whether the interactions went smoothly.

An example of a conversation with wh–type queries is presented in Table 1. The second column indicates the various types of utterances (dialogue acts) used by the system during this interaction. Note that during the slot-filling mode, the system has the initiative, whereas in the open-ended queries (cf. rows 15–22 in Table 1) the system gives the initiative to the participant and only produces feedback responses. Table 2 presents a conversation with yes–no queries following the conversation in Table 1. In a conversation with only yes–no queries, a wh– query may be presented if a slot has been observed with *no-value* (cf. rows 9–10 in Table 1) or when the participant disputes the system's current best estimate (cf. rows 6–9 in Table 2).

S	Introduction	Could you help me with some details about this view?
P		Yes
S	wh– query	Do you see a building on the far left?
P		Yes
S	Backchannel	Mh-hmm
S	wh– query	What color does the building have?
P		Cream color [Green color]
S	Reprise	Green
S	wh– query	How many floors is this building?
P		I don't know [I do no]
	…	
S	wh– query	Are there any important signs in this scene that I may have missed?
P		No
S	Backchannel	Uh-huh
S	wh– query	Is there anything peculiar in this scene that I should be familiar with
P		There is a bike painted on street just in front of me [is the right pink university in front of may]
S	Backchannel	Anything else?
P		Yeah its white paint [Yes light pink]
S	Backchannel	Mh-hmm
P		That's it [that that]
S	Backchannel	Ehm
P		I am done [okay]
S	Ending	Okay, thanks

Table 1: An example interaction with wh– queries for the visual scene in Figure 2. S: system, P: participant (here, u01 from Figure 5). ASR results are shown within square brackets.

S	Introduction	Could you help me with some details about this view?
P		Yes
S	yes–no query	Is the building on the far left visible to you?
P		Yes
S	Backchannel	Mh-hmm
S	yes–no query	Does this building have green color?
P		No
S	Backchannel	Uh-huh
S	wh– query	But than what color is the building?

P		white [light]
S	Repeat	Sorry, what color did you say?
P		white [light]
S	wh– query	How many stories is this building?
P		I don't know [I do rental]
S	Backchannel	Ehm
	…	
S	wh– query	Are there any important signs in this scene that I may have missed?
P		Ehm no [nine no]
S	Backchannel	Ehm
S	yes–no query	I am informed that there are other important landmarks or structures in this scene that I should be aware of, is it true?
P		No
S	Backchannel	uh-huh
S	Ending	Okay, thanks

Table 2: An example interaction with yes–no queries corresponding to the visual scene in Figure 2. S: system, P: participant (here u02 from Figure 5). ASR results are shown within square brackets.

5 Data analysis

We analyzed the data (15×8 interactions) collected from the experiment along the following tracks: first, we compare the majority value of the slots to the ground truth as given by a human annotator; second, we explore how the ground truth of slot-values could be estimated automatically; third, we also analyzed the instances where the participants disputed the system's current estimate of slot-values; and fourth, we examined the post-experimental questionnaires.

5.1 Rate of learning slot-values

A total of 197 slots were learned in the experiment. We analyzed how many slot-values had been correctly retrieved after 1, 2… 15 users. In Figure 6, 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 (according to the majority) and 96.4% of slot-values after interacting with the first six users. Another eight users, or fourteen in total, were required to learn all the slot-values correctly. The progression curve thus provides an estimate of how many users are required to achieve a specific percentage of slot-values correctly if majority is to be considered the ground truth. The curve "Not-in-Majority" indicates the number of slot with values that were not in the majority. Thus after interacting with the first user 20.8% of slot-values the system had obtained were not in majority and could be treated as incorrect. Note that the curves Majority and Not-in-Majority do not sum up to 100%, this is because we consider *no-value* as a valid slot-value, and treat the slot as unfilled. For example, 12.2% of the slots remained unfilled after interacting with the first user.

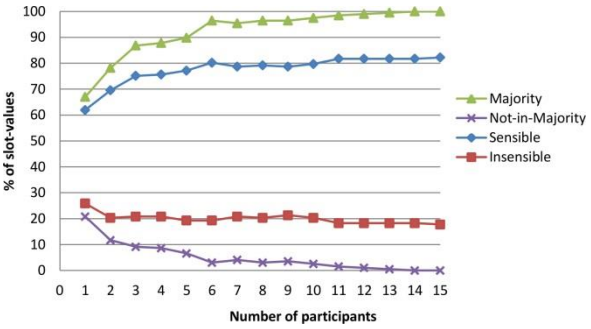

Figure 6: Rate of learning slot-values with two different estimates of ground truth

We also investigated how close the majority is to the actual truth. A human annotator (one of the coauthors) labeled all the obtained slot-values as either sensible or insensible, based on the combined knowledge from the corresponding maps, the visual scenes, and the set of obtained values. Thus a slot could have many sensible values. For example, various parts of a building could be painted in different colors. The progression curves "Sensible" and "Insensible" in Figure 6 illustrate the fraction of total slots for which the learned values were actually correct and incorrect, respectively. While the curve for sensible values follows the same pattern as the progression curve for majority as the estimate of ground truth, the percent of slot-values that were actually correct is always lower than the majority as ground truth, and it never reached 100%. The constant gap between the two curves suggests that some slot-values learned by the majority were not actually the ground truth. What led the majority into giving incorrect slot-values is left as a topic for future work.

As mentioned earlier, much of the slot-filling interaction involved buildings and their properties. Figure 7 illustrates that sensible values for most slots, pertaining to whether a building is visible, whether it is residential, whether it has shops, and the color of roof were obtained by interacting with only few participants. In contrast, properties such as color of the building and

number of stories required many more participants. This could be attributed to the fact that participants may have differences in perception about slot-values. As regards to whether there are signs on buildings, we observed that the recall is relatively low. This is largely due to lack of common ground among participants about what could be considered a sign. Our intentions with designing this prompt was to retrieve any peculiar detail on the building that is easy to locate: for us a sign suggesting a name of restaurant is as useful as the knowledge that the building has blue sunshade on the windows. Some participants understood this while other didn't.

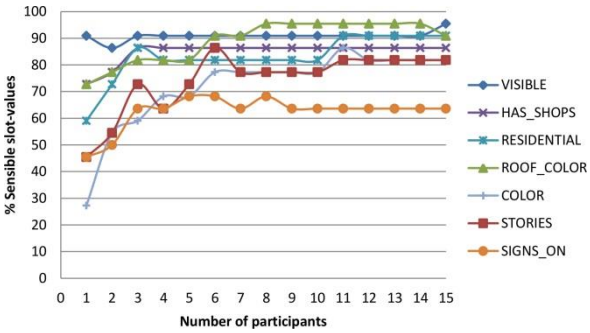

Figure 7: Learning rate of various slots for landmark type *building*

5.2 Estimated ground truth of slot-values

The 15 subjects in the in-lab experiment were all asked for the same information. In a real application, however, we want the system to only ask for slots for which it has insufficient or conflicting information. If the ground truth of a certain slot-value pair can be estimated with a certainty exceeding some threshold (given the quality requirements of the database, say 0.8), the system can consider the matter settled, and need not ask about that slot again. We therefore want to estimate the ground truth of slot-values along with a certainty measure. To this end, we use the CityCrowdSource Trust software package (Dickens & Lupu, 2014), which is based on the probabilistic approach for supervised learning when we have multiple annotators providing labels (possibly noisy) but no absolute gold standard, presented in Raykar et al. (2009).

Using this approach, a question concerning the color of a building, say with ID 24, (e.g. "What color is the building?") would be translated into several binary predicates COLOR_Red(24), COLOR_Brown(24), COLOR_Orange(24), etc. The justification for this binary encoding is that the different color values are not mutually exclusive: A building might of course have more than one color, and in many cases more than one color name might be appropriate even though the building has only one dominating color (e.g. to describe the color either as "brown" and "red" might be acceptable to most people). Figure 8 shows the incremental estimates for different colors for a certain building (OpenStreetMap ID 163966736) after 1, 2… 15 subjects had been asked. The answer from the first subject was erroneously recognized as "pink". The next 9 subjects all referred to the building as "brown". Among the final subjects, 3 subjects referred to building as "red", and 2 subjects as "brown". The final truth estimates are 0.98 for "brown", 0.002 for "red", and 0.00005 for "pink". The diagram shows that if the certainty threshold is set to 0.8, the value "brown" would have been established already after 4 subjects.

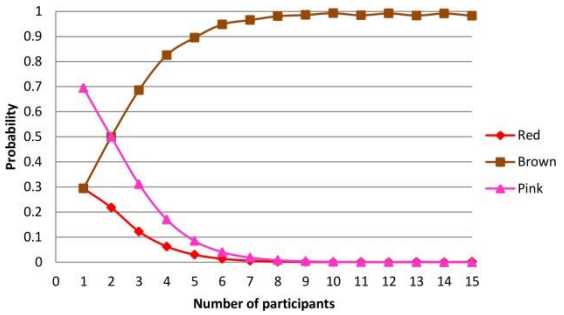

Figure 8: Probabilities of different estimated ground truth values for the color of a certain building

5.3 Disputed slot-values

We also examined all system questions of yes–no type that received negative answers, i.e. instances where the participants disputed the system's current best estimate (based on majority vote) of a slot-value. Among the 95 such instances, the system's current best estimate was actually insensible only on 43 occasions. In 30 of these instances the participants provided a rectified slot-value that was sensible. For the remaining 13 instances the new slot-values proposed by the participant were actually insensible. There were 52 instances of false disputations, i.e. the system's current estimate of a slot-value was sensible, but the participants disputed it. 6 of these occurrences were due to errors in speech recognition, but for the remaining 46 occasions, error in grounding the intended landmark (15), users' perception of slot-values (3), and ambiguity in what the annotator terms as sensible slot-values (28), (e.g. whether there are signs on a building (as discussed in Section 5.1)) were iden-

tified as the main reasons. This suggests that slots (i.e. attributes) that are often disputed may not be easily understood by users.

5.4 Post-experimental questionnaire

As described above, the participants filled in a questionnaire after each interaction. They were asked to rate the system's familiarity with the visual scene based on the questions asked. A Mann–Whitney *U* test suggests that participants' perception of the system's familiarity with the visual scene was significantly higher for interactions with yes–no queries than interactions with wh– queries (U=1769.5, p= 0.007). This result has implications for the design choice for systems that provide as well as ask for information from users. For example, a pedestrian routing system can already be used to offer routing instructions as well as crowdsourcing information. The system is more likely to give an impression of familiarity with the surrounding, to the user, by asking yes–no type questions than wh–questions. This may influence a user's confidence or trust in using the routing system.

Since yes–no questions expect a "yes" or "no" in response, we therefore hypothesized that interactions with yes–no questions would be perceived smoother in comparison to interactions with wh– questions. However, a Mann–Whitney *U* test suggests that the participants perceived no significant difference between the two interaction types (U=1529.0, p= 0.248). Feedback comments from participants suggest that abrupt ending of open-ended interactions by the system (due to the simplistic model of detecting whether the user has anything more to say) gave users an impression that the system is not allowing them to speak.

6 Discussion and future work

We have presented a proof-of-concept study on using a spoken dialogue system for crowdsourcing street-level geographic information. To our knowledge, this is the first attempt at using spoken dialogue systems for crowdsourcing in this way. The system is fully automatic, in the sense that it (i) starts with minimal details – obtained from OpenStreetMap – about a visual scene, (ii) prompts users with wh– questions to obtain values for a predefined set of attributes; and (iii) assumes attribute-values with majority vote as its beliefs, and engages in yes–no questions with new participants to confirm them. In a data collection experiment, we have observed that after interacting with only 6 human participants the system acquires more than 80% of the slots with actually sensible values.

We have also shown that the majority vote (as perceived by the system) could also be incorrect. To mitigate this, we have explored the use of the CityCrowdSource Trust software package (Dickens & Lupu, 2014) for obtaining the probabilistic estimate of the ground truth of slot-values in a real crowd-sourcing system. However, it is important not only to consider the ground truth probabilities per se, but also on how many contributing users the estimate is based and the quality of information obtained. We will explore these two issues in future work.

We have observed that through open-ended prompts, the system could potentially collect a large amount of details about the visual scenes. Since we did not use any automatic interpretation of these answers, we transcribed key concepts in participants' speech in order to obtain an estimate of this. However, it is not obvious how to quantify the number of concepts. For example, we have learned that in Figure 2, at the junction ahead, there is: a *traffic-sign*, a *speed-limit* sign, a sign with *yellow* color, a sign with *red* color, a sign with *red boarder*, a sign that is *round*, a sign with some *text*, the *text* says *50*. These are details obtained in pieces from various participants. Looking at Figure 2 one can see that these pieces when put together refer to the speed-limit sign mounted on the traffic-signal at the junction. How to assimilate these pieces together into a unified concept is a task that we have left for future work.

Acknowledgement

We would like to thank the participants of the in-lab crowd-sourcing experiment. This work is supported by the EIT KIC project "*CityCrowdSource*", and the Swedish research council (VR) project *Incremental processing in multimodal conversational systems* (2011-6237).

Reference

Bartie, P. J., & Mackaness, W. A. (2006). Development of a Speech-Based Augmented Reality System to Support Exploration of Cityscape. *Transactions in GIS, 10*(1), 63-86.

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.

Dickens, L., & Lupu, E. (2014). *Trust service final deliverable report*. Technical Report, Imperial College, UK.

Haklay, M., & Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing, 7*(4), 12-18.

Janarthanam, S., Lemon, O., Liu, X., Bartie, P., Mackaness, W., Dalmas, T., & Goetze, J. (2012). Integrating Location, Visibility, and Question-Answering in a Spoken Dialogue System for Pedestrian City Exploration. In *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue* (pp. 134-136). Seoul, South Korea: Association for Computational Linguistics.

Krug, K., Mountain, D., & Phan, D. (2003). Webpark: Location-based services for mobile users in protected areas.. *GeoInformatics*, 26-29.

Parent, G., & Eskenazi, M. (2011). Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges. In *INTERSPEECH* (pp. 3037-3040). ISCA.

Raykar, V. C., Yu, S., Zhao, L. H., Jerebko, A., Florin, C., Valadez, G. H., Bogoni, L., & Moy, L. (2009). Supervised Learning from Multiple Experts: Whom to Trust when Everyone Lies a Bit. In *Proceedings of the 26th Annual International Conference on Machine Learning* (pp. 889-896). New York, NY, USA: ACM.

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 Berlin Heidelberg.

Skantze, G., & Al Moubayed, S. (2012). IrisTK: a statechart-based toolkit for multi-party face-to-face interaction. In *Proceedings of ICMI*. Santa Monica, CA.

Tom, A., & Denis, M. (2003). Referring to Landmark or Street Information in Route Directions: What Difference Does It Make?. In Kuhn, W., Worboys, M., & Timpf, S. (Eds.), *Spatial Information Theory. Foundations of Geographic Information Science* (pp. 362-374). Springer Berlin Heidelberg.

Appendix A

The table below lists slots (= landmark attributes) and the corresponding wh– and yes–no system questions. For attributes marked with * the dialogue manager switches to open-ended interaction mode.

Slot (=attribute)	System wh– questions	System yes–no questions
Visible: whether a particular landmark is visible from this view.	• Do you see a building on the far left? • Do you see another building in front of you? • Is there a junction on the right? • Do you see a traffic-signal ahead?	• Is the building on the far right visible to you? • I think there is another building in front of you, do you see it? • Can you see the junction on the right? • Are you able to see the traffic-signal ahead?
Color of the building	• What color does the building have? • What color is the building?	• I think this building is *red* in color, what do you think? • Does this building have *red* color?
Size of the building (in number of stories)	• How many floors do you think are there in this building • How many stories is this building	• I think there are *six* floors in this building, what do you think? • Is this building *six* storied?
Color of the building's *roof*	• What color does the roof of this building have? • What color is the roof of this building?	• I think the roof of this building is *orange* in color, what do you think? • Do you think that the roof of this building is *orange*?
Signs or ornamentation on the building	• Do you see any signs or decorations on this building?	• I think there is a sign or some decoration on this building, do you see it? • There may be a sign or a name on this building, do you see it?
Shops or restaurants in the building	• Are there any shops or restaurants in this building?	• I am informed that there are some shops or restaurants in this building, is it true? • I think there are some shops or restaurants in this building, what do you think?
Signs at landmarks	• Are there any important signs at the junction/crossing?	• I believe there is a sign at this junction/crossing, do you see it? • Do you see the sign at this junction/crossing?
**Description* of sign	• Could you describe this sign? • What does this sign look like? • Does the sign say something?	• Could you describe this sign? • What does this sign look like? • Does the sign say something?
**Signs* in the visual scene	• Are there any important signs in this scene that I may have missed? • Have I missed any relevant signs in this scene?	• There are some important signs in this scene that could be useful for my knowledge, am I right? • I am informed that there are some signs in this scene that are relevant for me, is it true?
**Landmarks* in the visual scene	• Are there any other important buildings or relevant structures in this scene that I should be aware of? • Is there anything particular in this scene that I should be familiar with? • Have I missed any relevant buildings or landmarks in this scene?	• I am informed that there are some important landmarks or structures in this scene that I should be aware of, is it true? • I have been told that there are some other things in this scene that I are relevant for me, is it true? • I believe I have missed some relevant landmarks in this scene, am I right?
**Description* of unknown *landmarks* e.g. shop, restaurant, building, etc.	• Could you describe it? • Could you describe them? • How do they look like?	• Could you describe it? • Could you describe them? • How do they look like?