Walk this way: Spatial grounding for city exploration[1]

Johan Boye[*], Morgan Fredriksson[†], Jana Götze[*],
Joakim Gustafson[*] and Jürgen Königsmann[†]

[*]KTH, School of Computer Science and Communication, 100 44 Stockholm, Sweden
[†]Liquid Media, Hammarby allé 34, 120 61Stockholm, Sweden

Abstract

Recently there has been an interest in spatially-aware systems for pedestrian routing and city exploration, due to the proliferation of smartphones with GPS receivers among the general public. Since GPS readings are noisy, giving good and well-timed route instructions to pedestrians is a challenging problem. This paper describes a spoken dialogue prototype for pedestrian navigation in Stockholm that addresses this problem by using various grounding strategies.

1 Introduction

Recent years have seen an immense proliferation of smartphones among the general public. Smartphones feature an open computing platform and GPS satellite tracking facilities, and coupled with geographic databases they allow for the creation of spatially-aware applications like routing a pedestrian from A to B (see e.g. Google Navigation (Google 2012a), or providing information of sites and services in the immediate vicinity (see e.g. Google Maps (Google 2012b)). Though some services and experimental systems rely on spoken *output* (Bartie et al. 2006, Krug et al. 2003), so far no such spatially-aware service has been based on spoken *dialogue* (e.g. the possibility for the user to intervene and ask "Should I turn left here?" or "What street am I walking on?" etc.). Furthermore, the advantage of the spoken-dialogue approach over a map-based approach is that many people find interpreting maps on a small screen to be strenuous and confusing (Looije et al, 2007). It is therefore safe to say that well-functioning spoken dialogue would be a valuable contribution to the plethora of spatially-aware mobile applications.

This paper describes an implemented dialogue system for helping a user explore the city of Stockholm. The system can either guide the user to a location of his choice ("I want to go to Odengatan"), or to specific spots chosen by the system, like a statue or an interesting architectural detail on a particular building. The latter setting in particular is interesting as it allows us to investigate various methods for producing referring spatial expressions, in order to help the user find quite small objects in a complex city environment.

In general, the city exploration problem addressed here is challenging since it involves the interpretation and generation of utterances within a rapidly changing spatial context under uncertainty.

2 Background

Many researchers within cognitive psychology have investigated how people give route instructions to one another (see e.g. Denis et al., 1999), and what the elements of a good route description are (see e.g. Lovelace et al. 1999, Tom and Denis, 2003). It is however not clear how these results transfer to computational models of route description generation. One finding is that a big portion of such dialogues are devoted to grounding; making sure that the dialogue partner actually sees and understands what is being referred to. Grounding is a well-studied phenomenon also in dialogue systems (see e.g. Traum 1999, Skantze 2007).

The implemented systems for guiding pedestrians have mostly been based on spoken output from the system, with little or no possibility for the user to provide information (Malaka and Zipf 2000, Krug et al. 2003, Jöst et al. 2005, Bartie et al 2006, Zipf and Jöst 2005). Spoken dialogue systems in spatial domains have mostly focused on non-dynamic contexts where the user can ask questions about a static map (e.g. Cai et al. 2003, Wang et al. 2008), on virtual environments such as computer games (e.g. Boye and Gustafson 2005, Boye et al. 2006, Skantze et al. 2006, Striegnitz et al 2011), in indoor environments (Cuayáhuitl et al 2011), or on natural-language interfaces to robots (Lemon et al. 2001, MacMahon et al. 2006, Johansson et al. 2011). Few if any researchers have so far addressed the topic of spoken natural-language dialogue with a user in a real, dynamic city environment.

3 Uncertainty and grounding

A recurring problem for any pedestrian routing system is to describe to the user how to get from his current position to the next node in the planned route. This has to be done reliably even though the user's position, speed and direction are uncertain due to possible errors in GPS readings. Giving simple instructions like "turn left here" is therefore a risky strategy; such instructions might be nonsensical for the user if he is not quite where the system believes him to be. Furthermore, the interpretation of left and right is not always clear, for instance in parks and open squares, or when the user is standing still without the system knowing which way he is facing. Therefore, before giving directions, it is often preferable that the system first *grounds* the user's current position and orientation by means of reference landmarks in the near vicinity.

Consider for instance the situation depicted in Figure 1. Here the system seeks to describe the route given by the route planner, first to node A, then (when the user reaches A) to B, then down a flight of stairs to node C, then turn left, etc. Before giving instructions, our system first calculates if there is a clear line of sight from the user's assumed position to a number of reference landmarks. It then se-

lects the most salient landmark, seeks to make the user aware of it, and describes the route relative to it. Here is a sample dialogue:

1. **System:** There is a fountain about 35 metres from here. Can you see it?
2. **User:** Yes.
3. **S:** Good! Please walk to the left of the fountain.
   (*user walks*)
4. **S:** Please turn right and walk to the top of the stairs.
5. **U:** What?
6. **S:** There is a flight of stairs leading down about 25 metres from here. Can you see it?
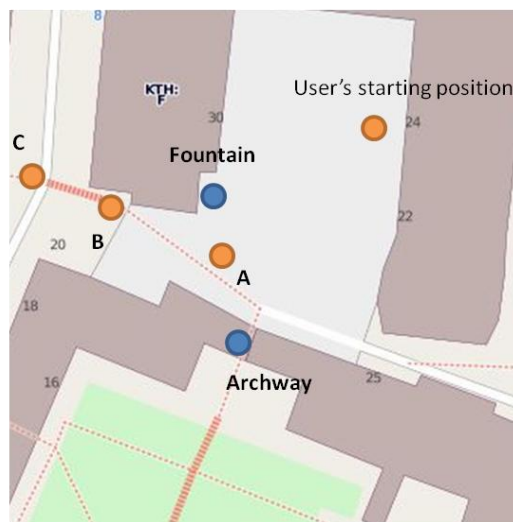

Figure 1: Generating route instructions.

In utterance 1, since there is no good way of describing node A, the system cannot ask directly about it. Instead, the system calculates that there are two describable landmarks visible from the user's presumed position; a fountain and an archway, of which the fountain is considered most salient. When the user confirms (utterance 2), the system gives the next instruction with a reference to the fountain. If the user had answered in the negative, the system would have proceeded to ask about another visible landmark. If all possibilities are exhausted, the user is asked to simply start walking, so the system can adjust his course if needed.

Determining salience and producing good referential expressions is a difficult problem in general. Salience measures used by our system include rarity (rare objects such as fountains are more salient than entrances to buildings), distance, uniqueness, and familiarity (objects that have been mentioned before in the dialogue are considered more salient, and are easily described, e.g. "the fountain that you passed before").

4 Uncertainty and re-planning

The system knows the user's position by means of the GPS receiver in the user's Android device. When GPS readings indicate that the user is within 20 metres of the next node in the planned route, the system issues the next instruction. Furthermore, the system can also use the GPS readings to estimate whether the user has misunderstood the latest instruction and is going off in the wrong direction. In the latter situation, the system will re-plan the route.

Unfortunately, the so-called "canyon effect" (Boriello et al. 2005) can introduce inaccuracies into GPS readings, and these errors can be quite substantial. Figure 2 shows a typical situation, in which the user is walking along the street (from left to right in the picture), and where the GPS readings (in red) are incorrect a large part of the time. These inaccuracies are a problem for two reasons.

Firstly, the user can appear to "miss" the 20-metre circle around the next node, and appear never to come sufficiently close. The result will be that no instruction is produced by the system at that node. Secondly, the user can appear to walk in the wrong direction when in fact he is not. Consider the situation 2 depicted in Figure 2 below. The user has passed the next node A, but GPS errors have prevented the system from registering this. At 2, the user is getting further and further away from A, and since the system is still considering A to be the next node, it appears as if the user is going the wrong way. Clearly, it would be very misleading and confusing for the user if the system would say "Please turn around" at this point.
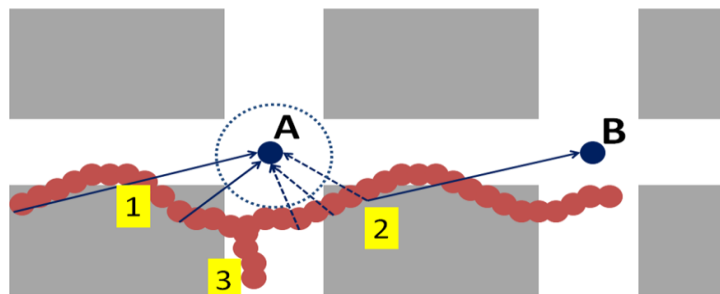

Figure 2: The user appears to "miss" the expected next node due to GPS drift.

The method we have adopted to address these problems is illustrated above. As long as the distance to the next node A is decreasing (situation 1), everything is fine. If the distance to the next node starts increasing (situation 2), the system checks the distance to the next-next node B as well. In situation 2, the distance to the next-next node B is *decreasing* while the distance to the expected next node A is *increasing*. If this pattern persists for 10 seconds, the system assumes that the user has passed the expected node A and is continuing in the correct direction.

Another possibility is when the distance is increasing both to the expected next node A and to the expected next-next node B (situation 3). If this pattern persists for 10 seconds, the system assumes that the user is walking in the wrong direction, and will issue re-planning.

5 Using visibility information

The system repeatedly performs visibility calculations to find out whether there is a free line-of-sight between two given points. Such visibility calculations are currently used for three purposes: Firstly, as mentioned in section 3, they are used to find candidates for referring expressions (the objects of which have to be visible from the user's assumed position). Secondly, they are used to produce better route plans. The system currently gets its data from OpenStreetMap (Haklay, 2008), and street objects in OpenStreetMap may contain many nodes very close to each other (in particular in roundabouts or curved streets). Consequently route plans can become very long. By iteratively weeding out any node visible from the preceding node, route plans become more suitable for narration. This process is depicted in Figure 3 below.
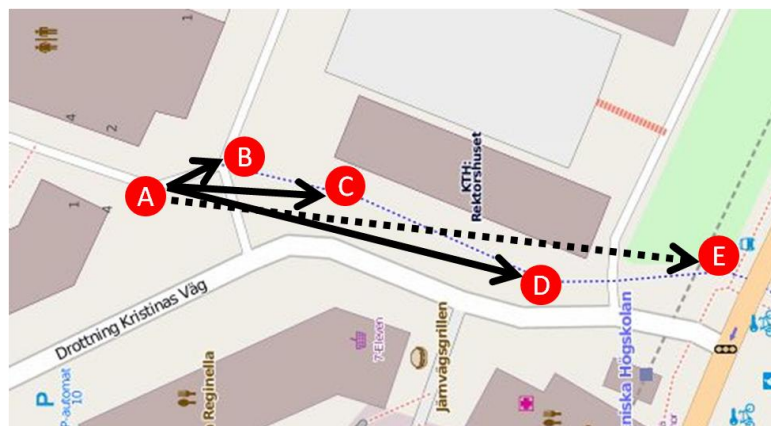

Figure 3: Weeding out nodes in a route plan using visibility information.

Here the produced route plan starts with the nodes A-E, in that order. The system checks visibility from A to B, from A to C, etc., and finds that D is the last visible node from A. It therefore concludes that B and C can be removed from the plan. The system then continues to check visibility from D to E, etc, until all unnecessary nodes are removed from the plan. However, no segment is allowed to be longer than 60 metres.

A third use of the visibility information is to produce better route instructions. The visibility calculations also return information on the streets, parks, etc, that were intersected by the visibility vector. This allows for instructions like "Now cross X street".

6 System architecture

The system is implemented to work speech-only and "eyes-free" – the user should not need to look down on a map on the screen, but rather be free to experience the city. The architecture described here is used both for the fully automatic system and for a Wizard-of-Oz data collection that preceded it. In the latter case, an operator GUI took the place of the Dialogue Manager. The operator GUI showed the user's position as a colored dot on a map, and used Google street view to show an approximation of the user's visual context.
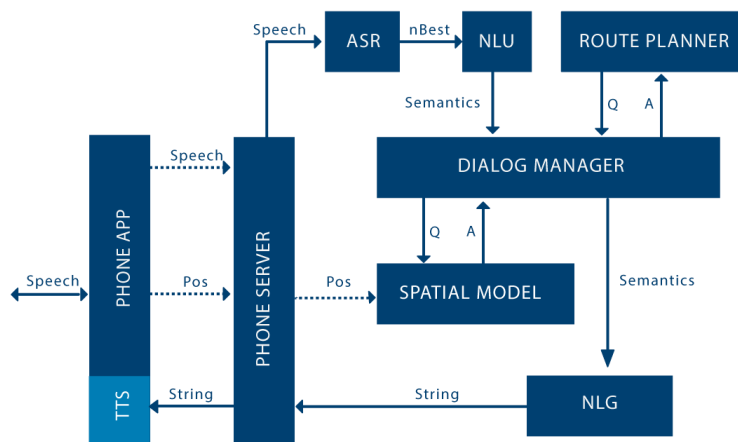

Figure 4: System architecture.

The user downloads a client app to his Android device which once started connects to a central phone server. The client app sends the sound stream from the microphone of the Android device to the phone server, and as soon as contact has been established with the GPS satellites, it also starts sending the coordinates (i.e. latitude-longitude pairs) of its current location. The sound stream is sent to speech recognition and parsing, and a semantic expression representing the utterance is sent to the Dialogue Manager (DM). Coordinates are sent directly to the DM. The dialogue manager updates its context model based on the input, and decides what to say to the user, and when to say it. The DM may also call an external planner to compute a route between two points in the city. For user speech input we are currently using a commercial off-the-shelf speech recognizer with a hand-written language model. For speech output, we use the built-in speech synthesizer on the Android device. The architecture also supports the use of server-side speech synthesis streamed to the handset, as well as the speech recognizer to be run on the handset. The latter feature would make it possible to maintain a dialogue in places where the 3G data connection fails.

Coordinates are also sent to the Spatial Model, which is a module that maintains the mapping from the logical representation of the city (in terms of buildings,

streets, etc.) to the algebraic representation (in terms of polygons, lines, and coordinates). The Spatial Model also performs the visibility calculations described in section 5. The polygon representation of the city is automatically generated from an export from OpenStreetMap, generated by indicating an area on the map. A minimal bounding rectangle is computed for each polygon in order to speed up visibility calculations, as it is faster to compute whether a line intersects a rectangle than an arbitrary polygon. If the Dialogue Manager needs to find out if B is visible from A, a request is sent to the Spatial Model, which first computes whether the line AB intersects any bounding rectangle in the entire city representation. If not, there is a clear line of sight from A to B. If the line intersects a bounding rectangle, a second more expensive calculation is carried out to check whether AB intersects the polygon inside the rectangle.

7 User experiment

In order to evaluate the strategies described in sections 3-5, we performed a user test with 8 subjects on 4 scenarios each. In each scenario, the user was guided to a specific spot in the city and asked to write down some inconspicuous detail (like the serial number on an electricity wiring box).

Table 1. User experiment

User #	Scenarios completed	#Instructions	Duration (mins)	# Re-plannings
1	4	49	28.2	3
2	4	59	34.6	10
3	4	77	40.4	5
4	4	68	28.3	10
5	1	6	2.1	0
6	4	60	27.7	4
7	4	82	35.6	14
8	4	48	24.9	0

As a rough estimate of the success of the implemented strategies, we note that 7 users managed to complete all 4 scenarios (1 user only completed 1 scenario due to technical problems), and that, on average, the system had to re-plan 1.6 times per completed scenario.

8 Concluding remarks

The system presented here routes pedestrians to their destination, using spoken dialogue to first ground reference landmarks used in the routing instructions. Ongoing work includes further user tests in a part of Stockholm in order to assess and improve the implemented strategies.

References

Bartie, P. and Mackaness, W (2006). Development of a speech-based augmented reality system to support exploration of cityscape. *Transactions in GIS*, 10(1):63–86.

Borriello, G., Chalmers, M., LaMarca, A. and Nixon, P. (2005) Delivering real-world ubiquitous location systems. *Communications of the ACM*, vol 8, issue 3, pp. 36-41.

Boye, J. and Gustafson, J. (2005) How to do dialogue in a fairy-tale world. *Proceedings of the 6th SIGDial workshop on discourse and dialogue,* Lisbon, Portugal.

Boye, J., Gustafson, J. and Wirén, M. (2006) Robust spoken language understanding in a computer game. *Journal of Speech Communication*, 48, pp. 335-353.

Cai, G. Wang, H. and MacEachren, A. (2003) Communicating Vague Spatial Concepts in Human-GIS Interactions: A Collaborative Dialogue Approach . In *Spatial Information Theory: Foundations of geographic information science,* LNCS Volume 2825/2003, pp 287-300.

Cuayáhuitl, H. and Dethlefs, N. (2011) Spatially-Aware Dialogue Control Using Hierarchical Reinforcement Learning. In *ACM Trans. on Speech and Language Processing (Special Issue on Machine Learning for Adaptive Spoken Dialogue Systems)*, vol. 7, no. 3, pp. 5:1-5:26

Denis, M., Pazzaglia, F., Cornoldi, C. and Bertolo, L. (1999) Spatial discourse and navigation: an analysis of route directions in the city of Venice. *Applied cognitive psychology,* vol 13, no 2.

Google Inc. (2012a) Google Maps Navigation.   http://www.google.com/mobile/navigation/. .

Google Inc. (2012b) Google Maps for mobile. http://www.google.com/mobile/maps.

Gustafson, J, Bell, L, Beskow, J, Boye, J, Carlson, R, Edlund, J, Granström, B, House, D and Wirén M (2000) AdApt – a multimodal conversational dialogue system in an apartment domain, *Proceedings of ICSLP 00.*

Haklay, M. (2008) OpenStreetMap: User-generated street maps. *Pervasive computing IEEE,* vol. 7, issue 4, pp. 12-18.

Johansson, M, Skantze, G, and Gustafson, J. (2011) Understanding Route Directions in Human-Robot Dialogue. *Proc SemDial 2011: Proceedings of the 15th Workshop on the Semantics and Pragmatics of Dialogue,* pages 19–27.

Jöst, M., Häussler, J., Merdes, M. and Malaka, R. (2005) Multimodal interaction for pedestrians: an evaluation study. In *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, pp 59–66.

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

Lemon, O., Bracy, A., Gruenstein, A., and Peters, S.(2001) A Multi-Modal Dialogue System for Human-Robot Conversation , In *Proc. NAACL.*

Looije, R., te Brake, G. and Neerincx, M. (2007) Usability engineering for mobile maps. In *Proceedings of Mobility'07, 4th int. conference on mobile technology, applications, and systems.*

Lovelace, K., Hegarty, M. and Montello, D. (1999) Elements of good route descriptions in familiar and unfamiliar environments. *Spatial Information Theory. Cognitive and computational foundations of geographic information science,* LNCS, 1661/1991, Springer-Verlag.

MacMahon, M., Stankiewicz, B. and Kuijpers, B. (2006) Walk the Talk: Connecting Language, Knowledge, Action in Route Instructions. *National Conf on Artificial Intelligence (AAAI-06).*

Malaka, R. and Zipf, A.. (2000) Deep map – challenging IT research in the framework of a tourist information system. *Information and communication technologies in tourism*, Springer.

Skantze, G. (2007). Making grounding decisions: Data-driven estimation of dialogue costs and confidence thresholds. In *Proceedings of SigDial* (pp. 206-210). Antwerp, Belgium.

Skantze, G., Edlund, J., & Carlson, R. (2006). Talking with Higgins: Research challenges in a spoken dialogue system. *Perception and Interactive Technologies*,pp. 193-196. Springer

Striegnitz, K., Denis, A., Gargett, A. Garoufi, K., Koller, A. and Theune, M. (2011) Report on the Second Challenge on Generating Instructions in Virtual Environments (GIVE-2.5). In *Proceedings of the Generation Challenges Session at the 13th European Workshop on Natural Language Generation (ENLG).*

Tom, A. and Denis, M. (2003) Referring to landmark or street information in route directions: What difference does it make?. *Spatial Information Theory: Foundations of geographic science.* LNCS 2825/2003, pp. 362-374. Springer-Verlag.

Traum, D. (1999) *Computational Models of Grounding in Collaborative Systems*. AAAI Technical Report FS-99-03

Wang, H., Cai, G. and MacEachren, A. (2008) GeoDialogue: A Software Agent Enabling Collaborative Dialogues between a User and a Conversational GIS . In *Tools with Artificial Intelligence, 2008. ICTAI '08. 20th IEEE.*

A Zipf and M. Jöst. (2005) Implementing adaptive mobile GI services based on ontologies - examples for pedestrian navigation support. *Computers, Environment and Urban Systems*, Special Issue on LBS and UbiGIS., 2005.