Human Evaluation of Conceptual Route Graphs for Interpreting Spoken Route Descriptions

Raveesh Meena, Gabriel Skantze and Joakim Gustafson

Department of Speech, Music and Hearing, KTH, Stockholm, Sweden
raveesh@csc.kth.se, {gabriel, jocke}@speech.kth.se

Abstract. We present a human evaluation of the usefulness of conceptual route graphs (CRGs) when it comes to route following using spoken route descriptions. We describe a method for data-driven semantic interpretation of route descriptions into CRGs. The comparable performances of human participants in sketching a route using the manually transcribed CRGs and the CRGs produced on speech recognized route descriptions indicate the robustness of our method in preserving the vital conceptual information required for route following despite speech recognition errors.

1 Introduction

It is desirable to endow urban robots with capabilities for engaging in spoken dialogue with passersby to seek route directions for autonomous navigation in unknown environments. Understanding spoken route descriptions mandates a robot's dialogue system to have a spoken language understanding (SLU) component that (i) is *robust* in handling automatic speech recognition (ASR) errors, (ii) learns *generalization* to deal with unseen concepts in free speech, and (iii) preserve the highly *structured relations* among various spatial and linguistic concepts present in route descriptions.

A SLU component in a dialogue system takes an ASR hypothesis as input and outputs a semantic representation that can be used by the dialogue manager to decide the next course of actions. A common way of representing navigational knowledge is the *route graph*. While varying level of details could be specified in a route graph (e.g. metric route graph), they are not representative of how humans structure information in route descriptions. Thus, a *conceptual route graph* (CRG) [1], is needed that can be used to represent human route descriptions semantically. In [2], we have presented a novel approach for data-driven semantic interpretation of manually transcribed route descriptions into CRGs. More recently, in [3] we applied this approach for semantic interpretation of spoken route descriptions. The results indicate that our approach is robust in handling ASR errors. The question as to whether the generated CRGs could actually be used by an agent in following the described route and arrive at the intended destination was left as future work.

In this paper, we evaluate the usefulness of the automatically extracted CRGs by asking human participants to sketch the described route on a map. Such an objective evaluation offers an alternative approach to evaluate our method: comparable human performances using the manually transcribed CRGs and the CRGs produced from speech recognized results would confirm the robustness of our method in preserving vital conceptual information for route following, despite speech recognition errors. In addition, a detailed analysis of human performances would help us (i) identify areas for further improvement in our method and the model, and also (ii) assess the usefulness of CRGs as a semantic representation for freely spoken route descriptions.

2 Previous work

It has been established in the literature that route descriptions contain a lot of redundant information whereas only a limited set of details are actually necessary for route following. These include descriptions about: the *landmarks* on the route, the *spatial relations*, the *controllers* that ensure traversal along the intended route, and the *actions* for changing orientation. Both data-driven and grammar based parsing approaches for semantic interpretation of route descriptions have been presented and evaluated for route following through human participants and/or robots in real and/or virtual environments [4-8]. Most of these works have focused on interpreting manually transcribed or human written route descriptions. Understanding verbal route descriptions has not received much attention. In [4] an ASR system has been used for recognizing verbal route descriptions, but the recognized text was translated to primitive routines using a translation scheme. In the following section, we briefly describe our data-driven approach for semantic interpretation of spoken route descriptions into CRGs, which have been shown to be useful in robot navigation [5].

2.1 A chunking parser for semantic interpretation

Our approach in [2] is a novel application of Abney's *chunking parser* [9], in which we apply the *Chunker* and the *Attacher* stages to automatically extract CRGs from route descriptions. A CRG is similar to a route graph in that nodes represent places where a change in direction takes place and edges connect these places. A route graph (or a *route*) may be divided into route *segments*, where each segment consists of an edge and an ending node where an action to change direction takes place. Conceptually, a segment consists of (i) *controllers* – a set of descriptions that guide the traversal along the edge, e.g. "*go straight down that road*", (ii) *routers* – a set of place descriptors that helps to identify the ending node, e.g. "*turn left at the post-office"*, and (iii) *action* – the action to take at the ending node in order to change direction. At least one of these three components is required in a route segment.

Fig. 1 illustrates an example CRG in which the nodes represent the semantic concepts and the edges their attributes. The concepts, their attributes and argument types are defined in the type hierarchy of the domain model using the specification in the JINDIGO dialogue framework [10].

To automatically extract CRGs, we first apply the *Chunker* stage of the Chunking parser for finding *base* concepts in a given sequence of words. Another chunk learner, namely the S*egmenter*, is then applied to automatically learn *route segments* in a sequence of base concepts. The *Attacher* takes a route segment as input and performs two tasks for each base concept present in it: First, it may assign a more specific concept class (like POSTOFFICE). To allow it to generalize, the Attacher also assigns all ancestor classes, based on the domain model (i.e. BUILDING for POSTOFFICE). The second task for the Attacher is to assign attributes, e.g. *direction*, and assign them values, e.g. →, which means that the interpreter should look for a matching argument in the right context. Table 1 illustrates these three stages for parsing the route description "*turn left at eh the post-office and then take*…"

Table 1. The three stages of the Chunking parser for interpreting route descriptions.

Chunker	[**ACTION** *turn*] [**DIRECTION** *left*] [**ROUTER** *at*] [**FP** *eh*] [**LANDMARK** *the post-office*] [**sCONT** *and then*] [**ACTION** *take*]
Segmenter	[**SEGMENT** [ACTION *turn*] [DIRECTION *left*] [ROUTER *at*] [FP *eh*] [LANDMARK *the post-office*]] [**SEGMENT** [sCONT *and then*] [ACTION *take*]]
Attacher	[SEGMENT [**TURN** (*direction*: →) *turn*] [**LEFT** *left*] [**AT** (*landmark*: →) *at*] [**DM** *eh*] [**POSTOFFICE** *the post-office*]] [SEGMENT [**DM** *and then*] [**TAKE** *take*]]

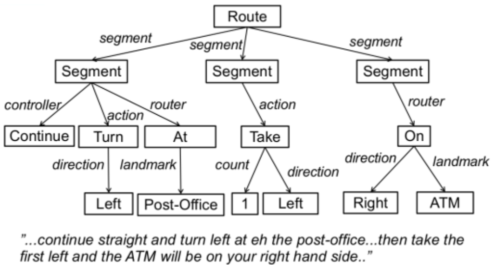

Fig. 1. An example Conceptual Route Graph.

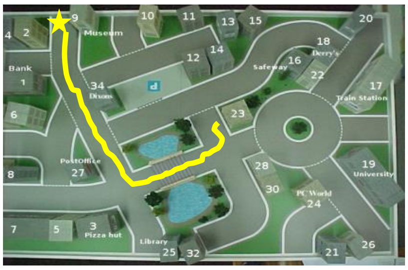

Fig. 2. The IBL map.

To measure the performance of our method we used the notion of Concept Error Rate (CER) – the weighted sum of the edits required in the manually transcribed CRG to obtain the extracted CRG. To evaluate our method we used the IBL corpora, which contain audio recordings and manual transcriptions of 144 spoken route instructions given in English [11]. Thirty five IBL transcriptions were manually annotated and used as the cross-validation set. Using the Linear Threshold Unit algorithms and best feature combinations discussed in [3], a baseline CER of 18.04 was obtained for comparing the Chunking parser's performance on speech recognized results.

Next, we trained an off-the-shelf ASR system with the remaining 108 route descriptions. For the best speech recognized hypothesis (mean WER = 27.59) for the route descriptions in the cross-validation set we obtained a CER of 28.15, i.e., a relative increase of mere 10.11 in CER. The relative increase in CER (R-CER) remains rather steady (*SD* = 2.80) with increase in WER. This illustrates the robustness of our method in dealing with speech recognition errors.

3 Method

Material: Six IBL route descriptions from the set of 35 were used for human evaluation. Care was taken in selecting routes to ensure that subjects could not guess the destination. For each route we obtained four instruction types: (1) the IBL manual transcription (ManTsc), (2) the manually annotated CRG (crgMAN), (3) the CRG extracted from the IBL manual transcription (crgCMT), and (4) the CRG extracted from the speech recognized route description (crgASR). The 24 items resulting from this combination were rearranged into four sets, each comprising of the six routes, but differing in the instruction type for the routes.

Subjects*:* A total of 16 humans (13 male and 3 female) participated in the evaluation. Participants ranged in the age from 16 to 46 (mean = 30.87, *SD* = 7.74). All, but one were researchers or graduate students in computer science.

Procedure: Participants were asked to sketch the route, on the IBL map (cf. Fig. 2, the star indicates the starting place), corresponding to the provided instruction. Each participant was individually introduced to the basic concept types in CRGs and shown how a route could be planned using the various nodes and sub-graphs in a CRG. Participants were asked to also mark concepts that they thought were absolutely necessary and strike-out what was redundant for the task at hand. Each of the four sets was evaluated by four different participants.

3.1 Results and analysis

We classified the 96 human performances under three categories: (1) FOUND: the participant arrived at the target building following the intended path, (2) ALM_THERE: the participant has almost arrived at the target building following the intended path, but did not correctly identify it among the others, (3) NOT_FOUND: the participant lost her way and did not arrive at the target building. Fig. 3 provides an overview of these performances across the four instruction types. One-way ANOVA test indicates a significant difference between only the human performances across crgASR and ManTsc instructions ($p < 0.05$). This is not surprising given that the crgASR instructions were produced from speech recognized results with WER of 47.64 (*SD* = 7.98) and have a R-CER of 27.35. However, there is no significant difference in performances across the crgMAN, crgCMT and crgASR instructions. This suggests that the conceptual information, required for human route following, present in Chunker parser produced CRGs is comparable with the information present in manually annotated CRGs, despite the CER of 20.29 and R-CER of 27.35 for crgCMT and crgASR instructions respectively.

These results confirm the robust performance of Chunking parser in dealing with speech recognition errors and preserving the vital conceptual information. Moreover, the results also suggest that improving the model (i.e. the CRG representation) to reduce the gap between human performances for ManTsc and crgMAN instructions will further enhance the human performances for Chunking parser extracted CRGs.

A closer analysis of the ALM_THERE (13) and NOT_FOUND (20) performances (a total of 33) suggest five general problem categories: (1) *SpatialR*: spatial relations, (2) *Controller*, (3) *Action*, (4) *Landmark*, and (5) *Other*: human errors. Across these five problem categories five sources were identified: (1) *Annotation*: an incorrect or underspecified manual annotation, (2) *ASR*: concepts insertion or deletion during speech recognition, (3) *ChunkingP*: Chunking parser errors, (4) *Model*: a limitation of the current model, and (5) *Human*: human judgments about the relevance or redundancy of a concept and executing actions.

The distribution of these error sources across the problem categories, as illustrated in Fig. 4, indicates that majority of the problems pertain to spatial relations (51.51%) and *Controller*s (24.24%). While some of the problems with the spatial relations are a result of incorrect and underspecified annotations (9.09%), which may have contributed to Chunking parser errors (9.09%) and to an extent to human judgments (21.21%), manual observations suggest that the overall human performance could have been better with the inclusion of additional spatial relation and *Controller* types in the model. We have refrained from elaborate annotations in the current model due to limited amount of training data. Human judgments were the source of half of the errors (51.51%). This indicates that it wasn't always easy to make the right decision about discarding or using concepts in the CRGs for route planning.

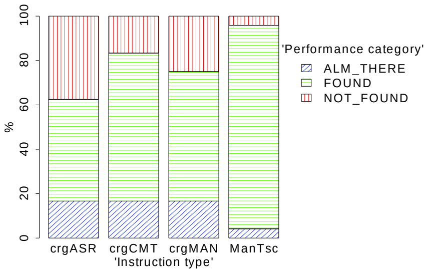

Fig. 3. Human performances across the instructions types.

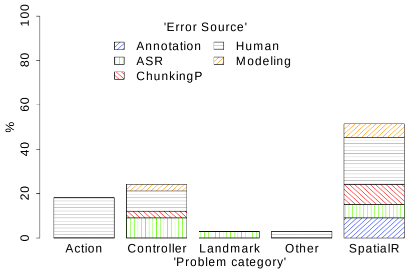

Fig. 4. Distribution of error sources across the problem categories.

4 Discussion and conclusion

From this human evaluation exercise we note that:

- *Controller*s with travel distance argument are vital for representing the extent of movement in a particular direction in route descriptions, such as "*follow the road to its end on the right is the treasure*" or "*a few buildings down from Pizza-Hut*".
- A requisite for proper grounding of the spatial relations in CRGs is resolving their *direction* or *landmark* arguments, or even both. The *Attacher's* role in attaching the concept RIGHT in CRG "[BUILDING *Tescos*] [AT (*landmark*: ⟵) *is on*] [RIGHT *right*]", as the *direction* argument for spatial relation AT is essential for locating the landmark.
- The CRG representations for spoken route description contain redundant concepts that arise from speech phenomena, such as pronominal references, anaphoric descriptions, self-repair and repetitions, about *landmarks* and *actions*. The CRG representation for "*you will take the third exit off…the third exit will be for Plymouth university…take this third exit*", contains two *actions* and four *landmarks*. Grounding this to a simple "*take the third exit*" would require additional approaches.
- ASR errors pose another challenge for an agent in route planning using the CRGs. Without access to the topological view of the environment a robot could not possibly infer erroneous concept insertions. To deal with this, we believe clarification or

reprise of route segments would be a prudent strategy, provided that the clarification sub-dialogue itself doesn't lead to further errors.

We have presented a human evaluation of the usefulness of conceptual route graphs – extracted from spoken route descriptions using our data-driven method – for route following. The comparable human performances on sketching the route using the manually transcribed and automatically extracted CRGs suggest no significant loss of conceptual information, required for route following, during the semantic interpretation of verbal route descriptions. This illustrates the robustness of our method in preserving vital conceptual information despite ASR errors. We observe that, extracting CRGs from spoken route descriptions mandates integration of approaches to counter speech phenomena, such as anaphoric descriptions and self-repairs, and using clarification strategies to recover from erroneous concept insertions during ASR.

Acknowledgement

This research was funded by the Swedish research council (VR) project "Incremental processing in multimodal conversational systems" (#2011-6237) and the European Commission project IURO, (#248314).

References

1. Müller, R., Röfer, T., Lankenau, A., Musto, A., Stein, K., & Eisenkolb, A. (2000). Coarse qualitative descriptions in robot navigation. In Freksa, C., Brauer, W., Habel, C., & Wender, K-F. (Eds.), *Spatial Cognition II* (pp. 265-276). Springer.
2. Johansson, M., Skantze, G., & Gustafson, J. (2011). Understanding route directions in human-robot dialogue. In *Proc. of SemDial* (pp. 19-27). Los Angeles, CA.
3. Meena, R., Skantze, G., & Gustafson, J. (2012). A Data-driven Approach to Understanding Spoken Route Directions in Human-Robot Dialogue. *Interspeech*. Portland, OR.
4. Bugmann, G., Klein, E., Lauria, S., & Kyriacou, T. (2004). Corpus-Based Robotics: A Route Instruction Example. In Groen, F. (Ed.), *IAS, vol. 8* (pp. 96-103).
5. Mandel, C., Frese, U., & Rofer, T. (2006). Robot navigation based on the mapping of coarse qualitative route descriptions to route graphs. In *Proc. of IEEE/RSJ IRS* (pp. 205-210). Beijing, China.
6. Kollar, T., Tellex, S., Roy, D., & Roy, N. (2010). Toward understanding natural language directions. In *Proc. of 5th ACM/IEEE HRI* (pp. 259-266). Piscataway, NJ, US. IEEE Press.
7. Pappu, A., & Rudnicky, A. I. (2012). The Structure and Generality of Spoken Route Instructions. In *Proc. of SIGDIAL* (pp. 99-107). Seoul, South Korea. ACL.
8. MacMahon, M., Stankiewicz, B., & Kuipers, B. (2006). Walk the talk: connecting language, knowledge, and action in route instructions. In *Proc. of the 21st national conference on Artificial intelligence - vol. 2* (pp. 1475-1482). AAAI Press.
9. Abney, S. (1991). Parsing by chunks. In Berwick, R. C., Abney, S. P., & Tenny, C. (Eds.), *Principle-Based Parsing: Computation and Psycholinguistics* (pp. 257-278). Kluwer.
10. Skantze, G. (2010). *Jindigo: a Java-based Framework for Incremental Dialogue Systems*. Technical Report, KTH, Stockholm, Sweden.
11. Kyriacou, T., Bugmann, G., & Lauria, S. (2005). Vision-based urban navigation procedures for verbally instructed robots. *Robotics and Autonomous Systems, 51*(1), (pp. 69-80).