Situated Language Understanding at 25 Miles per Hour

Teruhisa Misu, Antoine Raux*, Rakesh Gupta
Honda Research Institute USA
425 National Avenue
Mountain View, CA 94040
tmisu@hra.com

Ian Lane
Carnegie Mellon University
NASA Ames Research Park
Moffett Field, CA 93085

Abstract

In this paper, we address issues in situated language understanding in a rapidly changing environment – a moving car. Specifically, we propose methods for understanding user queries about specific target buildings in their surroundings. Unlike previous studies on physically situated interactions such as interaction with mobile robots, the task is very sensitive to timing because the spatial relation between the car and the target is changing while the user is speaking. We collected situated utterances from drivers using our research system, Townsurfer, which is embedded in a real vehicle. Based on this data, we analyze the timing of user queries, spatial relationships between the car and targets, head pose of the user, and linguistic cues. Optimized on the data, our algorithms improved the target identification rate by 24.1% absolute.

1 Introduction

Recent advances in sensing technologies have enabled researchers to explore applications that require a clear awareness of the systems' dynamic context and physical surroundings. Such applications include multi-participant conversation systems (Bohus and Horvitz, 2009) and human-robot interaction (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*.

We address yet another environment, where situated interactions takes place – a moving car. In the previous work, we collected over 60 hours of in-car human-human interactions, where drivers interact with an expert co-pilot sitting next to them in the vehicle (Cohen et al., 2014). One of the insights from the analysis on this corpus is that drivers frequently use referring expressions about their surroundings. (e.g. *What is that big building on the right?*) Based on this insight, we have developed Townsurfer (Lane et al., 2012; Misu et al., 2013), a situated in-car intelligent assistant. Using geo-location information, the system can answer user queries/questions that contain object references about points-of-interest (POIs) in their surroundings. We use driver (user) face orientation to understand their queries and provide the requested information about the POI they are looking at. We have previously demonstrated and evaluated the system in a simulated environment (Lane et al., 2012). In this paper, we evaluate its utility in real driving situations.

Compared to conventional situated dialog tasks, query understanding in our task is expected to be more time sensitive, due to the rapidly changing environment while driving. Typically, a car will move 10 meters in one second while driving at 25 mi/h. So timing can be a crucial factor. In addition, it is not well understood what kind of linguistic cues are naturally provided by drivers, and their contributions to situated language understanding in such an environment. To the best of our knowledge, this is the first study that tackles the issue of situated language understanding in rapidly moving vehicles.

In this paper, we first present an overview of the Townsurfer in-car spoken dialog system (Section 2). Based on our data collection using the system, we analyze user behavior while using the system focusing on language understanding (Section 3). Specifically, we answer the following research questions about the task and the system through data collection and analysis:

1. Is timing an important factor of situated language understanding?
2. Does head pose play an important role in language understanding? Or is spatial distance information enough?

* Currently with Lenovo.

Proceedings of the SIGDIAL 2014 Conference, pages 22–31,
Philadelphia, U.S.A., 18-20 June 2014.

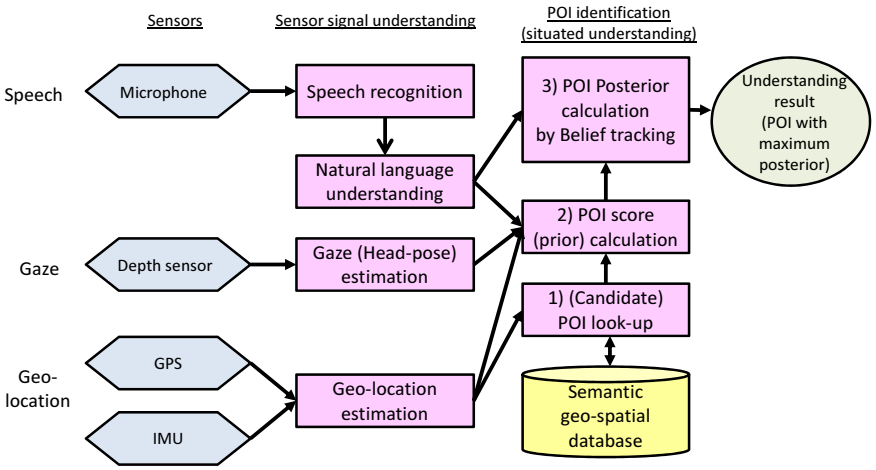

Figure 1: System overview of Townsurfer

Table 1: Example dialog with Townsurfer

U1:	What is *that place*. (POI in gaze)
S1:	This is Specialty Cafe, a mid-scale coffee shop that serves sandwiches.
U2:	What is *its* (POI in dialog history) rating.
S2:	The rating of Specialty Cafe is above average.
U3:	How about *that one* on the left. (POI located on the left)
S3:	This is Roger's Deli, a low-priced restaurant that serves American food.

3. What is the role of linguistic cues in this task? What kinds of linguistic cues do drivers naturally provide?

Based on the hypothesis obtained from the analysis for these questions, we propose methods to improve situated language understanding (Section 4), and analyze their contributions based on the collected data (Sections 5 and 6). We then clarify our research contributions through discussion (Section 7) and comparison with related studies (Section 8).

2 Architecture and Hardware of Townsurfer

The system uses three main input modalities, speech, geo-location, and head pose. Speech is the main input modality of the system. It is used to trigger interactions with the system. User speech is recognized, then requested concepts/values are extracted. Geo-location and head pose information are used to understand the target POI of the user query. An overview of the system with a process flow is illustrated in Figure 1 and an example dialog with the system is shown in Table 1. A video of an example dialog is also attached.

In this paper, we address issues in identifying user intended POI, which is a form of reference resolution using multi-modal information sources[1]. The POI identification process consists of the following three steps (cf. Figure 1). This is similar to but different from our previous work on landmark-based destination setting (Ma et al., 2012).

1) The system lists candidate POIs based on geo-location at the timing of a driver query. Relative positions of POIs to the car are also calculated based on geo-location and the heading of the car.

2). Based on spatial linguistic cues in the user utterance (e.g. *to my right, on the left*), a 2D scoring function is selected to identify areas where the target POI is likely to be. This function takes into account the position of the POI relative to the car, as well as driver head pose. Scores for all candidate POIs are calculated.

3) Posterior probabilities of each POI are calculated using the score of step 2 as prior, and non-spatial linguistic information (e.g. POI categories, building properties) as observations. This posterior calculation is computed using our Bayesian belief tracker called DPOT (Raux and Ma, 2011).

The details are explained in Section 4.

System hardware consists of a 3D depth sensor (Primesense Carmine 1.09), a USB GPS (BU-353S4), an IMU sensor (3DM-GX3-25) and a close talk microphone (plantronics Voyage Leg-

[1]We do not deal with issues in language understanding related to dialog history and query type. (e.g. General information request such as U1 vs request about specific property of POI such as U2 in Table 1)

end UC). These consumer grade sensors are installed in our Honda Pilot experiment car. We use Point Cloud Library (PCL) for the face direction estimation. Geo-location is estimated based on Extended Kalman filter-based algorithm using GPS and gyro information as input at 1.5 Hz. The system is implemented based on the Robot Operating System ROS (Quigley et al., 2009). Each component is implemented as a node of ROS, and communications between the nodes are performed using the standard message passing mechanisms in ROS.

3 Data Collection and Analysis

3.1 Collection Setting

We collected data using a test route. The route passes through **downtown** Mountain View[2] and **residential area** around Honda Research Institute. We manually constructed our database containing 250 POIs (businesses such as restaurants, companies) in this area. Each database entry (POI) has name, geo-location, category and property information explained in Section 3.4. POI geo-location is represented as a latitude-longitude pair (e.g. 37.4010,-122.0539). Size and shape of buildings are not taken into account. It takes about 30 minutes to drive the route. The major difference between residential area and downtown is the POI density. While each POI in downtown has on average 7.2 other POIs within 50 meters, in residential area POIs have only 1.9 neighbors. Speed limits also differ between the two (35 mi/h vs 25 mi/h).

We collected data from 14 subjects. They were asked to drive the test route and make queries about surrounding businesses. We showed a demo video[3] of the system to the users before starting the data collection. We also told them that the objective is a data collection for a situated spoken dialog system, rather than the evaluation of the whole system. We asked subjects to include the full description of the target POI within a single utterance to avoid queries whose understanding requires dialog history information[4]. Although the system answered based on the baseline strategy explained in Section 4.1, we asked subjects to ignore the system responses.

As a result, we collected 399 queries with a valid target POI. Queries about businesses that do not exist on our database (typically a vacant store) were excluded. The data contains 171 queries in downtown and 228 in residential area. The queries were transcribed and the user-intended POIs were manually annotated by confirming the intended target POI with the subjects after the data collection based on a video taken during the drive.

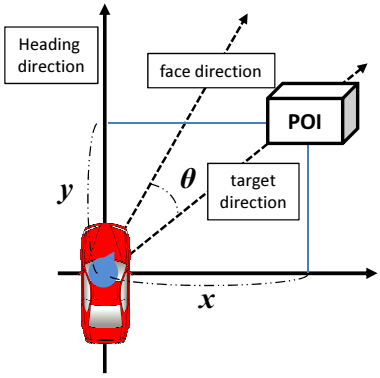

Figure 2: Parameters used to calculate POI score (prior)

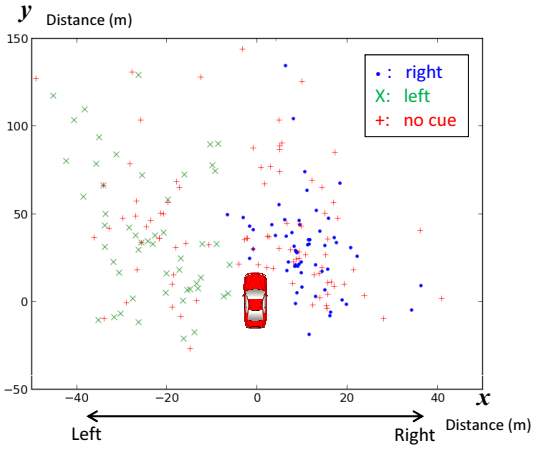

Figure 3: Target POI positions

3.2 Analysis of Spatial Relation of POI and Head Pose

We first analyze the spatial relation between position cues (right/left) and the position of the user-intended target POIs Out of the collected 399 queries, 237 (59.4%) of them contain either right or left position cue (e.g. *What is that on the left?*). The relation between the position cues (cf. Figure 2) and POI positions at start-of-speech timing [5] is plotted in Figure 3. The X-axis is a lateral distance (a distance in the direction orthogonal to the heading; a positive value means the right direction) and the Y-axis is an axial distance (a distance in the heading direction; a negative value means the POI is in back of the car.). The most obvious finding from the scatter plot is that right and left are pow-

[2]We assumed that a POI is in downtown when it is located within the rectangle by geo-location coordinates (37.3902, -122.0827) and (37.3954, -122.0760).

[3]not the attached one.

[4]Understanding including dialog history information is our future work.

[5]Specifically, the latest GPS and face direction information at that timing is used.

Table 2: Comparison of average and standard deviation of distance (in meter) of POI form the car

		ASR result timing		Start-of-speech timing	
Position cue	Site	Ave dist.	Std dist.	Ave dist.	Std dist.
Right/left	Downtown	17.5	31.0	31.9	28.3
	Residential	22.0	36.3	45.2	36.5
No right/left cue	Downtown	17.4	27.8	31.1	26.5
	Residential	38.3	45.9	52.3	43.4

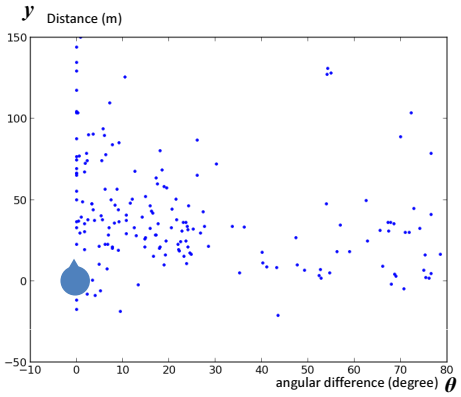

Figure 4: Relation between POI positions and head pose

erful cues for the system to identify target POIs. We can also see that the POI position distribution has a large standard deviation. This is partly because the route has multiple sites from downtown and residential area. Interestingly, while the average distance to the target POI in downtown is 37.0 meters, that of residential area is 57.4 meters.

We also analyze the relation between face direction and POI positions. Figure 4 plots the relation between the axial distance and the angular difference θ (between the user face direction and the target POI direction) (cf. Figure 2). The scatter plot suggests that the angular differences for distant target POIs is often small. For close target POIs the angular differences are larger and have a large variance[6].

3.3 Analysis of Timing

Referring expressions such as "the building on the right" must be resolved with respect to the context in which the user intended. However, in a moving car, such a context (i.e. the position of the car and the situation in the surroundings) can be very different between the time when the user starts speaking the sentence and the time they finish speaking it. Therefore, situated understanding must be very time sensitive.

To confirm and investigate this issue, we analyze the difference in the POI positions between the time the ASR result is output vs the time the user actually started speaking. The hypothesis is that the latter yields a more accurate context in which to interpret the user sentence. In contrast, our baseline system uses the more straightforward approach of resolving expressions using the context at the time of resolution, i.e. whenever the ASR/NLU has finished processing an utterance (hereafter "ASR results timing").

[6]We will discuss the reason for this in Section 6.2.

Table 3: User-provided linguistic cues

Category of linguistic cue	Percentage used (%)
Relative position to the car (right/left)	59.4
Business category (e.g. restaurant, cafe)	31.8
Color of the POI (e.g. green, yellow)	12.8
Cuisine (e.g. Chinese, Japanese, Mexican)	8.3
Equipments (e.g. awning, outside seating)	7.2
Relative position to the road (e.g. corner)	6.5

Specifically, we compare the average axial distance to the target POIs and its standard deviation between these two timings. Table 2 lists these figures broken down by position cue types and sites. The average axial distance from the car to the target POIs is often small at the ASR result timing, but the standard deviation is generally small at the start-of-speech timing. This indicates that the target POI positions at the start-of-speech timing is more consistent across users and sentence lengths than that at the ASR result timing. This result indicates the presence of a better POI likelihood function using the context (i.e. car position and orientation) at the start-of-speech timing than using the ASR result timing.

3.4 Analysis of Linguistic Cues

We then analyze the linguistic cues provided by the users. Here, we focus on objective and stable cues. We exclude subjective cues (e.g. *big, beautiful, colorful*) and cues that might change in a short period of time (e.g. *with a woman dressed in green in front*). We have categorized the linguistic cues used to describe the target POIs. Table 3 lists the cue types and the percentage of user utterances containing each cue type.

The cues that the users most often provided concern POI position related to the car (right and left). Nearly 60% of queries included this type of cue and every subject provided it at least once. The second most frequent cue is category of business, especially in downtown. Users also provided col-

ors of POIs. Other cues include cuisine, equipments, relative position to the road (e.g. *on the corner*).

Another interesting finding from the analysis is that the users provided more linguistic cues with increasing candidate POIs in their field of view. Actually, the users provided 1.51 categories in average per query in downtown, while they provided 1.03 categories in residential area. (cf. POI density in Section 3.2: 7.2 vs 1.9) This indicates that users provide cues considering environment complexity.

4 Methods for Situated Language Understanding

4.1 Baseline Strategy

We use our previous version (Misu et al., 2013) as the baseline system for situated language understanding. The baseline strategy consists of the following three paragraphs, which correspond to the process 1)-3) in Section 2 and Figure 1.

The system makes a POI look-up based on the geo-location information at the time ASR result is obtained. The search range of candidate POIs is within the range (relative geo-location of POIs against the car location) of -50 to 200 meters in the travelling direction and 100 meters to the left and 100 meters to the right in the lateral direction. The ASR result timing is also used to measure the distances to the candidate POIs.

POI priors are calculated based on the distance from the car (= axial distance) based on "the closer to the car the likely" principle. We use a likelihood function inversely proportional to the distance. We use position cues simply to remove POIs from a list of candidates. For example "right" position cue is used to remove candidate POIs that are located on < 0 position in the lateral distance. When no right/left cue is provided, POIs outside of 45 degrees from the face direction are removed from the list of candidates.

No linguistic cues except right/left are used to calculate POI posterior probabilities. So, the system selects the POI with the highest prior (POI score) as the language understanding result.

4.2 Strategies Toward Better Situated Language Understanding

To achieve better situated language understanding (POI identification) based on the findings of the analysis in Section 3, we modify steps 1)-3) as follows:

1. Using start-of-speech timing for the POI prior calculation

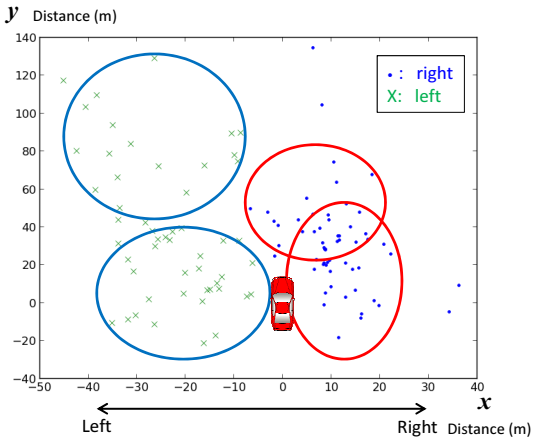

Figure 5: Example GMM fitting

2. Gaussian mixture model (GMM)-based POI probability (prior) calculation
3. Linguistic cues for the posterior calculation.

We use the start-of-speech timing instead of the time ASR result is output. Because the standard deviations of the POI distances are small (cf. Section 3.2), we expect that a better POI probability score estimation with the POI positions at this timing in the subsequent processes than the positions at the ASR result timing. The POI look-up range is the same as the baseline.

We apply Gaussian mixture model (GMM) with diagonal covariance matrices over the input parameter space. The POI probability (prior) is calculated based on these Gaussians. We use two input parameters of the lateral and axial distances for queries with right/left cue, and three parameters of the lateral and axial distances and the difference in degree between the target and head pose directions for queries without right/left cue. (The effect of the parameters is discussed later in Section 6.2.) We empirically set the number of Gaussian components to 2. An example GMM fitting to the POI positions for queries with right and left cues is illustrated in Figures 5. The center of ellipse is the mean of the Gaussian.

We use the five linguistic cue categories of Section 3.4 for the posterior calculation by the belief tracker. In the following experiments, we use either 1 or 0 as a likelihood of natural language understanding (NLU) observation. The likelihood for the category value is 1 if a user query (NLU result) contains the target value, otherwise 0. This corresponds to a strategy of simply removing candidate POIs that do not have the category values specified by the user. Here, we assume a clean POI database with all their properties annotated manually.

Table 4: Comparison of POI identification rate

Method	Success rate (%)
right/left linguistic cues, the-closer-the-likely likelihood, ASR result timing) (**Baseline**)	43.1
1) Start-of-speech timing	42.9
2) GMM-based likelihood	47.9
3) Linguistic cues	54.6
1) + 2)	50.6
1) + 3)	54.4
2) + 3)	62.2
1) + 2) + 3)	67.2

5 Experiments

We use manual transcriptions and natural language understanding results of the user queries to focus our evaluations on the issues listed in Section 1. We evaluate the situated language understanding (POI identification) performance based on cross validation. We use the data from 13 users to train GMM parameters and to define a set of possible linguistic values, and the data from the remaining user for evaluation. We train the model parameters of the GMM using the EM algorithm. Knowledge about the sites (downtown or residential area) is not used in the training[7].

We do not set a threshold for the presentation. We judge the system successfully understands a user query when the posterior of the target (user-intended) POI is the highest. The chance rate, given by the average of the inverse number of candidate POIs in the POI look-up is 10.0%.

6 Analysis of the Results

We first analyze the effect of our three methods described in Section 4.2. The results are listed in Table 4.

Simply using the POI positions at the start-of-speech timing instead of those of the ASR result timing did not lead to an improvement. This result is reasonable because the distances to target POIs are often smaller at the ASR result timing as we showed in Table 2. However, we achieved a better improvement (7.5% over the baseline) by combining it with the GMM-based likelihood calculation. The results supports our Section 3.3 hypothesis that the POI position is less dependent on users/scenes at the start-of-speech timing. The linguistic cues were the most powerful information for this task. The improvement over the baseline was 11.5%. By using these three methods together, we obtained more than additive improvement of 24.1% in the POI identification rate over the baseline[8]. The success rates per site were 60.8% in downtown and 71.9% in residential area.

[7]The performance was better when the knowledge was not used.

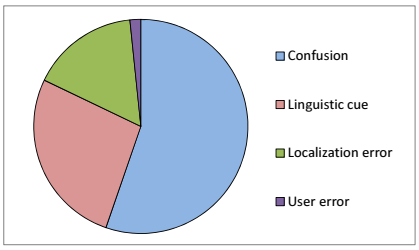

Figure 6: Breakdown of error causes

6.1 Error Analysis

To analyze the causes of the remaining errors, we have categorized the errors into the following four categories:

1. **Ambiguous references:** There were multiple POIs that matched the user query. (e.g. another *yellow building* sat next to the target)
2. **Linguistic cue:** The driver used undefined linguistic cues such subjective expressions or dynamic references objects (e.g. optometrist, across the street, colorful)
3. **Localization error:** Errors in estimating geo-location or heading of the car.
4. **User error:** There were errors in the user descriptions (e.g. user misunderstood the neighbor POI's outside seating as the target's)

The distribution of error causes is illustrated in Figure 6. More than half of the errors are due to reference ambiguity. These errors are expected to be resolved through clarification dialogs. (e.g. asking user *"Did you mean the one in front or back?"*) Linguistic errors might be partly resolved by using a better database with detailed category information. For dynamic references and subjective cues, use of image processing techniques will help. Localization errors can be solved by using high-quality GPS and IMU sensors. User errors were rare and only made in downtown.

6.2 Breakdown of Effect of the Spatial Distance and Head Pose

We then evaluate the features used for the POI prior calculation to investigate the effect of the input parameters of the lateral and axial distances

[8]For reference, the performances of "1) + 2) + 3)" were 62.9%, 67.2%, 66.1%, 67.2%, and 66.2% when the number of Gaussian components were 1, 2, 3, 4, and 5.

Table 5: Relation between the parameters used for the POI identification and success rates (%)

	query type	
parameters used	right/left	no cue
lateral (x) distance	58.6	51.2
axial (y) distance	59.5	53.7
face direction	43.3	44.4
lateral + axial ($x + y$)	73.8	54.3
lateral (x) + face direction	57.8	48.1
axial (y) + face direction	59.1	54.9
lateral + axial + face	68.4	57.4

and the difference in degree between the target and user face direction angles. Table 5 lists the relationship between the parameters used for the GMM-based likelihood calculation and the POI identification performances[9].

The results indicate that the axial distance is the most important parameter. We got a slight improvement by using the face direction information for the queries without right/left cue, but the improvement was not significant. On the other hand, use of face direction information for the right/left queries clearly degraded the POI identification performance. We think this is because the users finished looking at the POI and returned the face to the front when they started speaking, thus they explicitly provided right/left information to the system. However, we believe that using a long-term trajectory of the user face direction will contribute to an improve in the POI identification performance.

6.3 Breakdown of the Effect of Linguistic Cues

We then evaluate the effect of the linguistic cues per category. Table 6 lists the relationship between the categories used for the posterior calculation and the success rates. There is a strong correlation between the frequency of the cues used (cf. Table 3) and their contributions to the improvement in success rate. For example, business category information contributed the most, boosting the performance by 8.5%.

Another point we note is that the contribution of business category and cuisine categories is large. Because other categories (e.g. color) are not readily available in a public POI database (e.g. Google Places API, Yelp API), we can obtain reasonable performance without using a special database or image processing.

[9]Note that, we first determine the function to calculate POI scores (priors) based on the position cues, then calculate scores with the selected function.

Table 6: Effect of linguistic cues

linguistic cue category used	Success rate (%)
No linguistic cues (*)	50.6
(*) + Business category (e.g. cafe)	59.1
(*) + Color of the POI (e.g. green)	57.6
(*) + Cuisine (e.g. Chinese)	54.1
(*) + Equipments (e.g. awning)	53.9
(*) + Relative position (e.g. corner)	51.4

We also found that linguistic cues were especially effective in downtown. Actually, while the improvement[10] was 20.0% in downtown that for residential area was 14.4%. This mainly would be because the users provided more linguistic cues in downtown considering the difficulty of the task.

6.4 Using Speech Recognition Results

We evaluate the degradation by using automatic speech recognition (ASR) results. We use Google ASR[11] and Julius (Kawahara et al., 2004) speech recognition system with a language model trained from 38K example sentences generated from a grammar. An acoustic model trained from the WSJ speech corpus is used. Note that they are not necessarily the best system for this domain. Google ASR uses a general language model for dictation and Julius uses a mismatched acoustic model in terms of the noise condition.

The query success rate was 56.3% for Julius and 60.3% for Google ASR. We got ASR accuracies of 77.9% and 80.4% respectively. We believe the performance will improve when N-best hypotheses with confidence scores are used in the posterior calculating using the belief tracker.

7 Discussion

The main limitation of this work comes from the small amount of data that we were able to collect. It is not clear how the results obtained here would generalize to other sites, POI density, velocities and sensor performances. Also, results might depend on experimental conditions, such as weather, hour, season. Hyper-parameters such as the optimal number of Gaussian components might have to be adapted to different situations. We therefore acknowledge that the scenes we experimented are only a limited cases of daily driving activities.

[10]1) + 2) vs 1) + 2) + 3).

[11]Although it is not realistic to use cloud-based speech recognition system considering the current latency, we use this as a reference system.

However, the methods we propose are general and our findings should be verifiable without loss of generality by collecting more data and using more input parameters (e.g. velocity) for the POI prior calculation.

In addition, much future work remains to realize a natural interaction with the system, such as taking into account dialog history and selecting optimal system responses. On the other hand, we believe this is one of the best platform to investigate situated interactions. The major topics that we are going to tackle are:

1. Dialog strategy: Dialog strategy and system prompt generation for situated environments are important research topics, especially to clarify the target when there is ambiguity as mentioned in Section 6.1. The topic will include an adaptation of system utterances (entrainment) to the user (Hu et al., 2014).
2. Eye tracker: Although we believe head pose is good enough to estimate user intentions because we are trained to move the head in driving schools to look around to confirm safety, we would like to confirm the difference in this task between face direction and eye-gaze.
3. POI identification using face direction trajectory: Our analysis showed that the use of face direction sometimes degrades the POI identification performance. However, we believe that using a trajectory of face direction will change the result.
4. Database: We assumed a clean and perfect database but we are going to evaluate the performance when noisy database is used. (e.g. A database based on image recognition results or user dialog log.)
5. Feedback: Koller et al. (2012) demonstrated referential resolution is enhanced by giving gaze information feedback to the user. We would like to analyze the effect of feedback with an automotive augmented reality environment using our 3D head-up display (Ng-Thow-Hing et al., 2013).

8 Related Work

The related studies include a landmark-based navigation that handles landmarks as information for a dialog. Similar system concepts have been provided for pedestrian navigation situations (Janarthanam et al., 2013; Hu et al., 2014), they do not handle a rapidly changing environment.

Several works have used timing to enhance natural interaction with systems. Rose and Horvitz (2003) and Raux and Eskenazi (2009) used timing information to detect user barge-ins. Studies on incremental speech understanding and generation (Skantze and Hjalmarsson, 2010; Dethlefs et al., 2012) have proved that real-time feedback actions have potential benefits for users. Komatani et al. (2012) used user speech timing against user's previous and system's utterances to understand the intentions of user utterances. While the above studies have handled timing focusing on (para-)linguistic aspect, our work handles timing issues in relation to the user's physical surroundings.

Recent advancements in gaze and face direction estimation have led to better user behavior understanding. There are a number of studies that have analyzed relationship between gaze and user intention, such as user focus (Yonetani et al., 2010), preference (Kayama et al., 2010), and reference expression understanding (Koller et al., 2012), between gaze and turn-taking (Jokinen et al., 2010; Kawahara, 2012). Nakano et al. (2013) used face direction for addressee identification. The previous studies most related to ours are reference resolution methods by Chai and Prasov (2010), Iida et al. (2011) and Kennington et al. (2013). They confirmed that the system's reference resolution performance is enhanced by taking the user's eye fixation into account. However, their results are not directly applied to an interaction in a rapidly changing environment while driving, where eye fixations are unusual activities.

Marge and Rudnicky (2010) analyzed the effect of space and distance for spatial language understanding for a human-robot communication. Our task differs with this because we handle a rapidly changing environment. We believe we can improve our understanding performance based on their findings.

9 Conclusion

We addressed situated language understanding in a moving car. We focused on issues in understanding user language of timing, spatial distance, and linguistic cues. Based on the analysis of the collected user utterances, we proposed methods of using start-of-speech timing for the POI prior calculation, GMM-based POI probability (prior) calculation, and linguistic cues for the posterior calculation to improve the accuracy of situated language understanding. The effectiveness of the proposed methods was confirmed by achieving a significant improvement in a POI identification task.

10 Acknowledgments

The authors would like to thank Yi Ma at Ohio State University for his contributions to the development of HRItk.

References

D. Bohus and E. Horvitz. 2009. Models for Multiparty Engagement in Open-World Dialog. In *Proc. SIGDIAL*, pages 225–234.

J. Chai and Z. Prasov. 2010. Fusing eye gaze with speech recognition hypotheses to resolve exophoric reference in situated dialogue. In *Proc. EMNLP*.

D. Cohen, A. Chandrashekaran, I. Lane, and A. Raux. 2014. The hri-cmu corpus of situated in-car interactions. In *Proc. IWSDS*, pages 201–212.

N. Dethlefs, H. Hastie, V. Rieser, and O. Lemon. 2012. Optimising incremental dialogue decisions using information density for interactive systems. In *Proc. EMNLP*, pages 82–93.

Z. Hu, G. Halberg, C. Jimenez, and M. Walker. 2014. Entrainment in pedestrian direction giving: How many kinds of entrainment? In *Proc. IWSDS*, pages 90–101.

R. Iida, M. Yasuhara, and T. Tokunaga. 2011. Multimodal reference resolution in situated dialogue by integrating linguistic and extra-linguistic clues. In *Proc. IJCNLP*, pages 84–92.

S. Janarthanam, O. Lemon, X. Liu, P. Bartie, W. Mackaness, and T. Dalmas. 2013. A multithreaded conversational interface for pedestrian navigation and question answering. In *Proc. SIGDIAL*, pages 151–153.

K. Jokinen, M. Nishida, and S. Yamamoto. 2010. On eye-gaze and turn-taking. In *Proc. EGIHMI*.

T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano. 2004. Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository. In *Proc. ICSLP*, volume IV.

T. Kawahara. 2012. Multi-modal sensing and analysis of poster conversations toward smart posterboard. In *Proc. SIGDIAL*.

K. Kayama, A. Kobayashi, E. Mizukami, T. Misu, H. Kashioka, H. Kawai, and S. Nakamura. 2010. Spoken Dialog System on Plasma Display Panel Estimating User's Interest by Image Processing. In *Proc. 1st International Workshop on Human-Centric Interfaces for Ambient Intelligence (HCIAmi)*.

C. Kennington, S. Kousidis, and D. Schlangen. 2013. Interpreting situated dialogue utterances: an update model that uses speech, gaze, and gesture information. In *Proc. SIGDIAL*.

A. Koller, K. Garoufi, M. Staudte, and M. Crocker. 2012. Enhancing referential success by tracking hearer gaze. In *Proc. SIGDIAL*, pages 30–39.

K. Komatani, A. Hirano, and M. Nakano. 2012. Detecting system-directed utterances using dialogue-level features. In *Proc. Interspeech*.

I. Lane, Y. Ma, and A. Raux. 2012. AIDAS - Immersive Interaction within Vehicles. In *Proc. SLT*.

Y. Ma, A. Raux, D. Ramachandran, and R. Gupta. 2012. Landmark-based location belief tracking in a spoken dialog system. In *Proc. SIGDIAL*, pages 169–178.

M. Marge and A. Rudnicky. 2010. Comparing Spoken Language Route Instructions for Robots across Environment Representations. In *Proc. SIGDIAL*, pages 157–164.

T. Misu, A. Raux, I. Lane, J. Devassy, and R. Gupta. 2013. Situated multi-modal dialog system in vehicles. In *Proc. Gaze in Multimodal Interaction*, pages 25–28.

Y. Nakano, N. Baba, H. Huang, and Y. Hayashi. 2013. Implementation and evaluation of a multimodal addressee identification mechanism for multiparty conversation systems. In *Proc. ICMI*, pages 35–42.

V. Ng-Thow-Hing, K. Bark, L. Beckwith, C. Tran, R. Bhandari, and S. Sridhar. 2013. User-centered perspectives for automotive augmented reality. In *Proc. ISMAR*.

M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Ng. 2009. ROS: an open-source Robot Operating System. In *Proc. ICRA Workshop on Open Source Software*.

A. Raux and M. Eskenazi. 2009. A Finite-state Turn-taking Model for Spoken Dialog Systems. In *Proc. HLT/NAACL*, pages 629–637.

A. Raux and Y. Ma. 2011. Efficient probabilistic tracking of user goal and dialog history for spoken dialog systems. In *Proc. Interspeech*, pages 801–804.

R. Rose and H. Kim. 2003. A hybrid barge-in procedure for more reliable turn-taking in human-machine dialog systems. In *Proc. Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 198–203.

G. Skantze and A. Hjalmarsson. 2010. Towards incremental speech generation in dialogue systems. In *Proc. SIGDIAL*, pages 1–8.

K. Sugiura, N. Iwahashi, H. Kawai, and S. Nakamura. 2011. Situated spoken dialogue with robots using active learning. *Advance Robotics*, 25(17):2207–2232.

Table 7: Example user utterances

- What is that blue restaurant on the right?
- How about this building to my right with outside seating?
- What is that Chinese restaurant on the left?
- Orange building to my right.
- What kind of the restaurant is that on the corner?
- The building on my right at the corner of the street.
- What about the building on my right with woman with a jacket in front
- Do you know how good is this restaurant to the left?
- Townsurfer, there is an interesting bakery what is that?
- Is this restaurant on the right any good?

S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy. 2011. Understanding natural language commands for robotic navigation and mobile manipulation. In *Proc. AAAI*.

R. Yonetani, H. Kawashima, T. Hirayama, and T. Matsuyama. 2010. Gaze probing: Event-based estimation of objects being focused on. In *Proc. ICPR*, pages 101–104.

11 Appendix

Test route:
`https://www.google.com/maps/preview/dir/Honda+Research+Institute,+425+National+Ave+%23100,+Mountain+View,+CA+94043/37.4009909,-122.0518957/37.4052337,-122.0565795/37.3973374,-122.0595982/37.4004787,-122.0730021/Wells+Fargo/37.4001639,-122.0729708/37.3959193,-122.0539449/37.4009821,-122.0540093/@37.3999836,-122.0792529,14z/data=!4m21!4m20!1m5!1m1!1s0x808fb713c225003d:0xcf989a0bb230e5c0!2m2!1d-122.054006!2d37.401016!1m0!1m0!1m0!1m0!1m5!1m1!1s0x0:0x86ca9ba8a2f15150!2m2!1d-122.082546!2d37.388722!1m0!1m0!1m0!3e0`