Landmark-based Location Belief Tracking in a Spoken Dialog System

Yi Ma
The Ohio State University
Columbus, OH 43210
may@cse.ohio-state.edu

Antoine Raux, Deepak Ramachandran, Rakesh Gupta
Honda Research Institute, USA
425 National Ave, Mountain View, CA 94043
{araux,dramachandran,
rgupta}@hra.com

Abstract

Many modern spoken dialog systems use probabilistic graphical models to update their belief over the concepts under discussion, increasing robustness in the face of noisy input. However, such models are ill-suited to probabilistic reasoning about spatial relationships between entities. In particular, a car navigation system that infers users' intended destination using nearby landmarks as descriptions must be able to use distance measures as a factor in inference. In this paper, we describe a belief tracking system for a location identification task that combines a semantic belief tracker for categorical concepts based on the DPOT framework (Raux and Ma, 2011) with a kernel density estimator that incorporates landmark evidence from multiple turns and landmark hypotheses, into a posterior probability over candidate locations. We evaluate our approach on a corpus of destination setting dialogs and show that it significantly outperforms a deterministic baseline.

1 Introduction

Mobile devices such as smart phones and in-car infotainment systems have generated demand for a new generation of location-based services such as local business search, turn-by-turn navigation, and social event recommendation. Accessing such services in a timely manner through speech is a crucial requirement, particularly on the go when the user is unable to resort to other modalities e.g. where safety regulations prohibit drivers from using buttons or a touchscreeen while driving.

In such systems, a Point of Interest (POI) or a destination such as a restaurant, store or a public place is often specified. For example, a car navigation system needs the user to input the destination before giving directions. Similarly, a photo tagging application must allow its users to designate the location where a picture was taken. While postal addresses can be used to unambigously identify locations, they are often either unknown or hard for users to remember. A more natural (though potentially ambiguous) means of specifying locations is to use *landmarks* such as "`the Italian restaurant near Red Rock cafe on Castro Street`" or "`the bakery near that mall with a Subway and a 7 Eleven`". A location-based dialog system that understands referring expressions using landmarks could lead to more succinct dialogs, higher recognition accuracy and a greater appearance of intelligence to the user.

We present a system that performs *belief tracking* over multiple turns of user speech input to infer the most probable target location. The user interacts with the system through speech in order to specify a target location, and may include references to one or more landmarks. Such a system must handle two sources of uncertainty. First, ASR is notoriously error-prone and modern ASR engines provide ranked lists of possible interpretations of speech input rather than single hypotheses. Second, the suitability of a particular landmark or its likelihood of usage by the speaker depends on a number of factors such as distance, size and prominence of the landmark, familiarity of the user and his expectation of

Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL), pages 169–178,
Seoul, South Korea, 5-6 July 2012. ©2012 Association for Computational Linguistics

common ground for understanding. These factors, or at least the resulting variability, must be taken into account when making inferences about target locations from landmark-based expressions.

The first source of ambiguity (speech understanding) has been the target of research on belief tracking (Mehta et al., 2010; Raux and Ma, 2011; Thomson and Young, 2010). In previous work, the concepts of interest are entities that are ontologically related (i.e. with *is-a* or *has-a* relations), thus discrete probabilistic graphical models such as DBNs have generally sufficed as representations. But these models are ill-suited for dense continuous spatial relations like the distance between any two locations on a map. In this paper, we introduce a *kernel-based belief tracker* as a probabilistic model for inferring target locations from (uncertain) landmarks. The kernel-based representation allows a natural way to weigh the suitability of a landmark and the speech understanding confidence. The output of this tracker is combined with that of a Dynamic Probabilistic Ontology Tree (DPOT) (Raux and Ma, 2011), which performs ontological reasoning over other features of the target location, to give a posterior distribution over the intended location. We evaluate our approach on a new corpus of location setting dialogs specially collected for this work and find it to significantly outperform a deterministic baseline.

2 Related Work

In the context of a location-based dialog system, Seltzer et al. (2007) describes a speech understanding system designed to recognize street intersections and map them to a database of valid intersections using information retrieval techniques. Robustness is achieved by exploiting both words and phonetic information at retrieval time, allowing a soft-matching of the ASR result to the canonical intersection name. Their approach is specifically targeted at intersections, to the exclusion of other types of landmarks. While intersections are frequently used as landmarks in North America (where their study was conducted), this is not always the case in other cultures, such as Japan (Suzuki and Wakabayashi, 2005), where points of interests such as train stations are more commonly used. Also, their approach, which is framed as speech understanding, does not exploit information from previous dialog turns to infer user intention.

Landmarks have been integrated in route directions (Pierre-emmanuel Michon, 2001; Tversky and Lee, 1999) with significant use at origin, destination and decision points. Further, landmarks have been found to work better than street signs in wayfinding (Tom and Denis, 2003). The multimodal system described in (Gruenstein and Seneff, 2007) supports the use of landmarks from a limited set that the user specifies by pointing at the map and typing landmark names. While this allows the landmarks (and their designations) to be of any kind, the burden of defining them is on the user.

Spatial language, including landmarks, has also been the focus of research within the context of human-robot interaction. (Huang et al., 2010; MacMahon et al., 2006) describe systems that translate natural language directions into motion paths or physical actions. These works focus on understanding the structure of (potentially complex) spatial language and mapping it into a representation of the environment. Issues such as imperfect spoken language understanding have not been investigated in this context. Similarly, this vein of spatial language research has traditionally been conducted on small artificial worlds with a few dozen objects and places at most, whereas real-world location-based services deal with thousands or millions of entities.

3 Hybrid Semantic / Location Belief Tracking

Our belief tracking system consists of two trackers running in parallel: a DPOT belief tracker (Raux and Ma, 2011) and a novel kernel-based location tracker. The final inference of user intentions is produced by combining information from the two trackers. The general idea is to rerank the user goals given spatial information provided by the location tracker.

3.1 Semantic Belief Tracker

We perform belief tracking over non-landmark concepts such as business name and street using a Dynamic Probabilistic Ontology Tree (DPOT) (Raux and Ma, 2011). A DPOT is a Bayesian Network composed of a tree-shaped subnetwork representing the (static) user goal (*Goal Network*), connected to

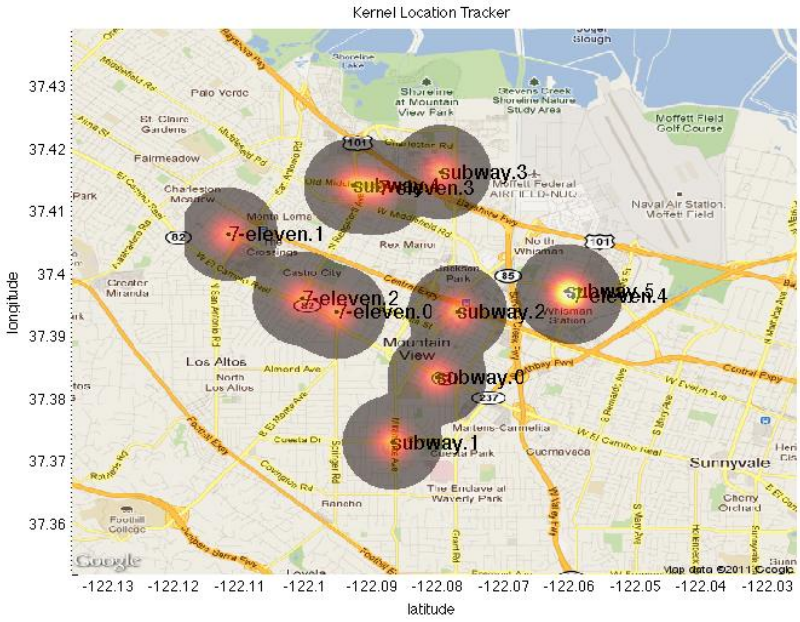

Figure 1: Top view heat map of spatial distribution with landmarks `Subway` and `7 Eleven` over potential target places in Mountain View, CA

a series of subnetworks representing the evidence gathered from each successive dialog turn (*Evidence Networks*). Details of the model and an efficient inference method for posterior probability computations can be found in (Raux and Ma, 2011).

In the context of this paper, the purpose of the semantic tracker is to update a list of the most likely target locations using attributes of that location provided by the user (see Figure 2). In a local business database, such attributes include `Business Name`, `Street`, `Category` (e.g. Japanese restaurant or convenience store), etc. The structure and parameters of the Goal Network encode probabilistic ontological relations between the attributes (e.g. a Mcdonalds would be described as a fast-food restaurant with high probability) that can be exploited during inference. These can be derived from expert knowledge, learned from data, or as is the case in our experimental system, populated from a database of local businesses (see section 4). After each user utterance, the DPOT outputs a ranked list of user goal hypotheses (an example goal hypothesis is [`Category=italian restaurant,Street=castro street`]). Each hypothesis is converted into a query to the backend database, and the posterior probability of the hypothesis is split equally among all matching entries. This results in a ranked list of database entries corresponding to the system's belief over potential target locations, with potentially many entries having the same probability.

3.2 Kernel-based Location Tracker

Landmark concepts extracted by the Natural Language Understanding module (NLU) are passed to the location tracker, which maintains a distribution over coordinates of potential target locations. Each such landmark concept is treated as evidence of spatial proximity of the target to the landmark and the distribution is accordingly updated. Any location in the database can serve as a landmark observation, including major POIs such as train stations or public facilities. If the name of a generic chain store with multiple locations such as `Subway` is used for the landmark, then an observation corresponding to each individual location is added to the tracker.

For each observed landmark ℓ, the location tracker constructs a 2-dimensional Gaussian kernel with mean equal to the longitude and latitude of the landmark ($\mu_\ell = (long_\ell, lat_\ell)$) and a fixed covari-

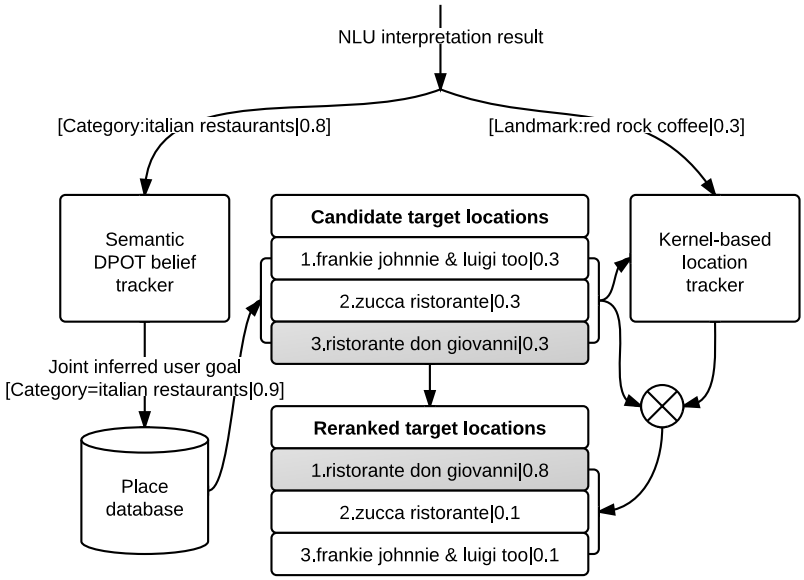

Figure 2: Overview of the hybrid semantic / location belief tracking approach; the database entry in shade is the underlying true target place to which the provided landmark is close

ance matrix Σ_ℓ for each landmark:

$$\mathbf{\Phi}_\ell(\boldsymbol{t}) = \frac{1}{2\pi|\Sigma|^{1/2}}\exp(-\frac{1}{2}(\boldsymbol{t}-\mu_\ell)^T\Sigma_\ell^{-1}(\boldsymbol{t}-\mu_\ell))$$

This kernel density determines the conditional probability that the target is at coordinates $\boldsymbol{t} = (long_t, lat_t)$ given the fixed landmark ℓ. The covariance matrix Σ_ℓ and hence the shape of the kernel can be adjusted for different landmarks depending on considerations such as the familiarity, size and prominence of the landmark (a large historic monument is likely to be used as a landmark for locations much further away than a small corner grocery store) etc.

The probability density of the location $\boldsymbol{t}$ being the target is then given by a weighted mixture model:

$$Pr(\boldsymbol{t}|L) = \sum_{\ell \in L} w_\ell \mathbf{\Phi}_\ell(\boldsymbol{t}) \quad (1)$$

where L is the set of candidate landmarks returned by the NLU (see Section 4.1) up to the current turn and w_ℓ is set to the confidence score of ℓ from the NLU. Thus candidate landmarks that have higher confidence in the NLU will contribute more strongly to the total likelihood. Since $Pr(\boldsymbol{t}|L)$ is a density function, it is unnormalized. In Figure 1, we show the kernel tracker distribution for a dialog state where `Subway` and `7 Eleven` are provided as landmarks.

The kernel density estimator is a simple approach to probabilistic spatial reasoning. It is easy to implement and requires only a moderate amount of tuning. It naturally models evidence from multiple speech hypotheses and multiple provided landmarks, and it benefits from accumulated evidence across dialog turns. It can also potentially be used to model more general kinds of spatial expressions by using appropriate kernel functions. For example, '`Along Castro street`' can be modeled by a Gaussian with an asymmetric covariance matrix such that the shape of the resulting distribution is elongated and concentrated on the street. While '`Two blocks away from ...`' could be modeled by adding an extra "negative" density kernel that extends from

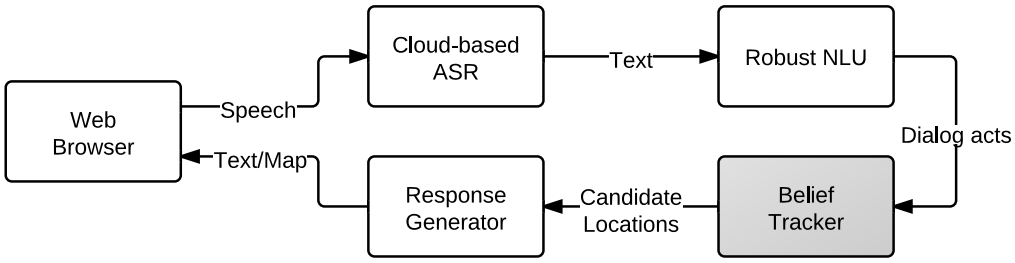

Figure 3: Overview of the Destination Setting System

the center of the landmark to a distance two blocks away.

3.3 Combining the Two Trackers

At each turn, the updated results from the Semantic and Location tracker are combined to give a single ranked list of likely target locations. In Figure 2, this process is illustrated for a dialog turn where two possible concepts are identified – a category attribute [`Category:italian restaurant`] and a landmark [`Landmark:red rock coffee company`]. These are passed to the DPOT tracker and the location tracker respectively. The output of the DPOT is used to retrieve and score matching database entries. The score for each entry is reweighted by the kernel density estimator measured at the coordinates of the location [1]:

$$Pr(e_{ij}) \quad = \quad (\frac{p_i}{N_i})^{\nu} \times Pr(e_{ij}|L) \qquad (2)$$

where N_i is the number of matching database entries retrieved from ith goal hypothesis (having joint probability p_i) and e_{ij} is the j^{th} such entry ($j \in [1..N_i]$). The exponent ν for the posterior term is introduced to account for scale difference between the semantic score and the kernel density.

The set of candidate entries can then be reranked according to Eq 2 and returned as the output of the combined belief tracker.

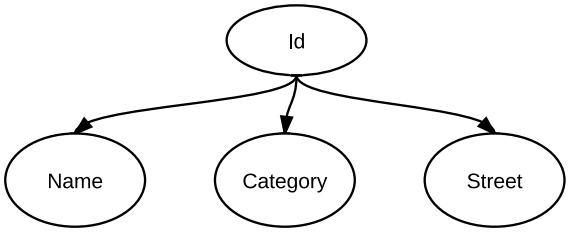

Figure 4: Structure of the Goal Network for the experimental system.

4 Evaluation

4.1 Experimental System

The architecture of our experimental system is shown in Figure 3. The web client, shown in Figure 5, runs in the participant's web browser and displays the target location of the current scenario using the Google Map API. The user's goal is to convey this target location to the system through speech only.

The system backend consists of a database of 2902 businesses located in Mountain View, California with their name, street, street number, business category, latitude and longitude provided. The grammar rules for the NLU and the probability tables in the DPOT are populated from this database.

The web client captures the user speech and sends it to our server with a push-to-talk interface based on the WAMI toolkit (Gruenstein et al., 2008). The server uses a commercial cloud-based ASR service with generic acoustic and language models, which were not adapted to our task. The n-best list of hypotheses from the ASR is sent to our robust natural

[1]The scores are renormalized to between 0 and1.

language understanding module for parsing.

Our NLU uses a hybrid approach combining a weighted finite-state transducer (WFST) with string matching based rescoring of the output. The WFST incorporates out-of-grammar word loops that allow skipping input words at certain points in the parse[2]. This parser robustly maps free form utterances (e.g. "`Okay let's go to that Italian place near, uh..., Red Rock Cafe, on Castro`") to semantic frames (e.g. [`Category=italian restaurant, Street=castro street, Landmark=red rock coffee company`]).

The NLU confidence score is computed based on the number of words skipped while parsing, and how close the important concept words match the canonical phrases found in the database. For instance, "`Red Rock Cafe`" matches the canonical name "`Red Rock Coffee Company`" with high confidence because rare words (`Red`, `Rock`) are identical, and differing but common words (`Cafe`, `Coffee`, `Company`) have a low weight in the score. The string matching score is based on the term-frequency/inverse document frequency (TF-IDF) metric commonly used in information retrieval. In our case, the weight of different terms (IDF) is estimated based on their frequency of occurrence in different database entries (i.e. how uniquely they describe a matching entry). We use the secondstring open-source library (Cohen et al., 2003) for string matching. For any ASR hypothesis, the NLU is likely to generate several parses which are all merged in a global list of candidate parses.

For each candidate parse, the system generates a set of dialog acts (one per concept in the parse) which are input to the belief tracker with their confidence score. Following the approach described in section 3, dialog acts corresponding to the Landmark concept are sent to the kernel-based location belief tracker, while all other concepts are sent to a Dynamic Probabilistic Ontology Trees (DPOT) semantic belief tracker, whose structure is shown in Figure 4. We use a two-level tree. The value of the root node (Id) is never directly observed and represents the database entry targeted by the user. The leaf nodes correspond to the relevant attributes `Name`, `Category`, and `Street`. For any database entry e, attribute a and value of that attribute v_a, the conditional probability $P(a = v_a|Id = e)$ is set to 1 if the value of a is v_a for entry e in the database, and to 0 otherwise. For attributes such as `Category`, which allow several possible values for each entry, the probability is split equally among valid values. After each user utterance, the network is augmented with a new Evidence Network capturing the possible interpretations and their likelihood, as computed by the NLU. The posterior probability distribution over user goals is computed and rescored using the kernel-based location tracker.

Finally, the Response Generator takes the highest scoring target location from the belief tracker and sends it back to the web client which displays it on the map and also indicates what are the values of the `Name`, `Category`, and `Street` concepts for the top belief (see Figure 5). If the top belief location does not match the goal of the scenario, the user can speak again to refine or correct the system belief. After the user has spoken 5 utterances, they also get the choice of moving on to the next scenario (in which case the dialog is considered a failure).

4.2 Data collection

To evaluate our approach, we ran a data collection experiment using the Amazon Mechanical Turk online marketplace. We defined 20 scenarios grouped into 4 Human Intelligence Tasks (HITs). Figure 5 shows a screen shot of the web interface to the system. In each scenario, the worker is given a target location to describe by referring to nearby landmark information. The target locations were chosen so as to cover a variety of business categories and nearby landmarks. The compensation for completing each set of 5 scenarios is 1 US dollar. Before their first scenario, workers are shown a video explaining the goal of the task and how to use the interface, in which they are specifically encouraged to use landmarks in their descriptions.

At the beginning of each scenario, the target location is displayed on the map with a callout containing a short description using either a generic category (e.g. `Italian restaurant`, `Convenience store`) or the name of a chain store (e.g. `Subway`, `Mcdonalds`). The worker

[2]This module is implemented using the OpenFST library (Allauzen et al., 2007)

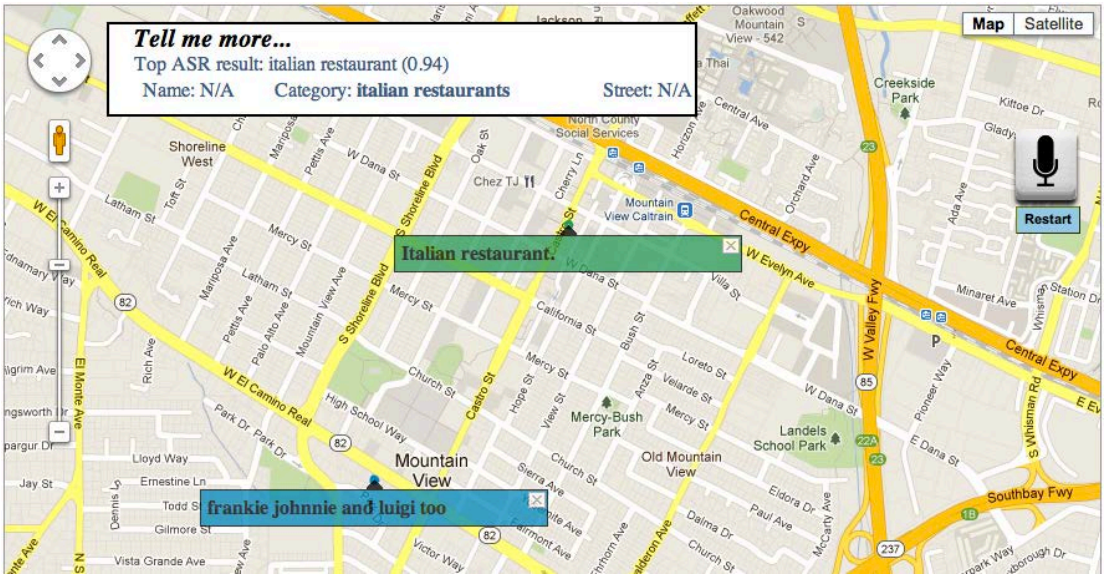

Figure 5: Screen capture of the data collection web interface where the target location is an Italian restaurant (in green, underlying target place is [`Ristorante Don Giovanni`]) and after the first turn user input '`Italian restaurant`' with a system belief [`Frankie, Johnnie & Luigi, Too`] in blue returned without any landmark information provided so far

then interacts with the system described in section 4.1 until either the system's top belief matches the target location, or they decide to skip the scenario.

4.3 Data Statistics

Overall, 99 workers participated in the data collection, providing 948 dialogs (2,869 utterances, 3 turns per scenario on average), which two of the authors manually transcribed and annotated for dialog acts. 76% of the dialogs (46% of utterances) contained a reference to a landmark. Other strategies commonly used by workers to uniquely identify a location include using a category or chain name and a street, as well as explicitly mentioning the target business name (although workers were explicitly discouraged form doing so). Figure 7 in appendix provides one example dialog from the corpus.

Overall, the workers provided 203 unique landmarks, of which 143 (70%) are in the database.

Workers were able to set the target destination within 5 turns in 60.1% of the dialogs, which we hereafter refer to as task successes. However, based on the manual transcripts, 19.0% of the dialogs could not have succeeded with the current system because the workers used landmark or attributes that do not appear in the database. Since the focus of this study is robustness rather than coverage, we base our evaluation on the remaining 768 dialogs, which we split between a development set of 74 dialogs and a test set of 694 dialogs. On this test set, the live system has a task success rate of 70.6%. By inspecting the log files, we noticed that runtime issues such as timeouts prevented the system from getting any belief from the belief tracker in 6.3% of the dialogs.

The mean Word Error Rate (WER) per worker on the test set is 27.5%. There was significant variability across workers, with a standard deviation 20.7%. Besides the usual factors such as acoustic noise and non-native accents, many of the errors came from the misrecognition of business names, due to the fact that ASR uses an open-ended language model that is tuned neither to Mountain View, nor to businesses, nor to the kind of utterances that our set up tends to yield, which is a realistic situation for large scale practical applications.

Concept precision of the top scoring NLU hypothesis is 73.0% and recall is 57.7%. However, when considering the full list of NLU hypotheses and using an oracle to select the best one for each turn, precision increases to 89.3% and recall to 66.2%, underscoring the potential of using multiple input hypotheses in the belief tracker.

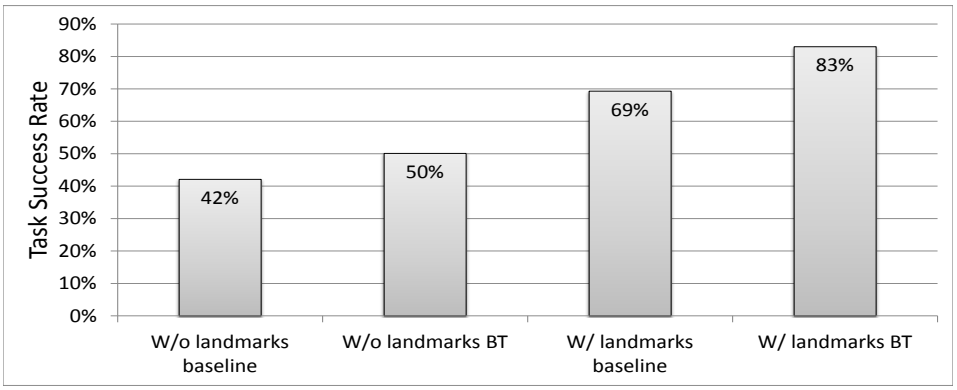

Figure 6: Batch evaluation of the proposed (BT) and baseline approaches with and without landmark information.

4.4 Batch Results

To further analyze the performance of our approach, we conducted a series of batch experiments on the data collected with the runtime system. We first tuned the parameters of the belief tracker ν and Σ_l (see section 3) on the development set ($\nu = 3$ and Σ_l corresponds to a circular Gaussian with standard deviation 500 meters).

We compare the tuned proposed belief tracking system (labeled BT) with three other versions. First, we define a deterministic baseline system which, at each turn, updates its belief by overwriting each concept's value with the value found in the top NLU hypothesis. Based on this (single) user goal hypothesis, we query the database to retrieve matching entries. If the current goal hypothesis contains a `Landmark` concept, the baseline system selects the matching entry that is closest to any location matching the landmark name, by computing the pairwise distance between candidate target locations and landmarks.

We also compute the performance of both the baseline and our proposed approach without using landmark information at all. In these versions, the belief over the attributes (`Name`, `Street`, and `Category`) is updated according to either the top NLU hypothesis (baseline) or the DPOT model (BT) and the first matching database entry is returned, ignoring any landmark information.

Figure 6 shows the task success of each of the four versions on the test set. First, it is clear that landmark information is critical to complete the tasks in this corpus since both systems ignoring landmarks perform significantly worse than their counterparts. Second, the belief tracking approach significantly outperforms the deterministic baseline (83.0% vs 69.3%, $p < 0.001$ using sign test for matched pairs).

To further analyze the performance of the system in different input conditions, we split the dialogs based on their measured concept accuracy (expressed in terms of concept F-measure). All dialogs with an F-measure higher than the median (70.0%) are labeled as high-accuracy, while the other half of the data is labeled as low-accuracy. While both the proposed approach and the baseline perform similarly well for high-accuracy dialogs (task success of resp. 96.0% and 92.8%, difference is not statistically significant), the difference is much larger for low-accuracy dialogs (70.0% vs 45.8%, $p < 0.001$) confirming the robustness of the landmark-based belief tracking approach when confronted with poor input conditions.

5 Conclusion

In this paper, we have explored the possibilities of incorporating spatial information into belief tracking in spoken dialog systems. We proposed a landmark-based location tracker which can be combined with a semantic belief tracker to output inferred joint user goal. Based on the results obtained from our batch experiments, we conclude that integrating spatial information into a location-based dialog system could improve the overall accuracy of belief tracking significantly.

References

Cyril Allauzen, Michael Riley, Johan Schalkwyk, Wojciech Skut, and Mehryar Mohri. 2007. Openfst: A general and efficient weighted finite-state transducer library. In *Proceedings of the Ninth International Conference on Implementation and Application of Automata (CIAA) Lecture Notes in Computer Science*, volume 4783, pages 11–23. Springer.

W.W. Cohen, P. Ravikumar, and S.E. Fienberg. 2003. A comparison of string distance metrics for name-matching tasks. In *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, pages 73–78.

Alexander Gruenstein and Stephanie Seneff. 2007. Releasing a multimodal dialogue system into the wild: User support mechanisms. In *In Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 111–119, September.

A. Gruenstein, I. McGraw, and I. Badr. 2008. The wami toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces. In *Proceedings of the 10th international conference on Multimodal interfaces*, pages 141–148. ACM.

Albert Huang, Stefanie Tellex, Abe Bachrach, Thomas Kollar, Deb Roy, and Nick Roy. 2010. Natural language command of an autonomous micro-air vehicle. In *International Conference on Intelligent Robots and Systems (IROS)*.

M. MacMahon, B. Stankiewicz, and B. Kuipers. 2006. Walk the talk: Connecting language, knowledge, and action in route instructions. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1475. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

N. Mehta, R. Gupta, A. Raux, D. Ramachandran, and S. Krawczyk. 2010. Probabilistic ontology trees for belief tracking in dialog systems. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 37–46. Association for Computational Linguistics.

Michel Denis Pierre-emmanuel Michon. 2001. When and why are visual landmarks used in giving directions? In D. R. Montello, editor, *Spatial Information Theory, Volume 2205 of Lecture Notes in Computer Science*, pages 292–305. Springer, Berlin.

A. Raux and Y. Ma. 2011. Efficient probabilistic tracking of user goal and dialog history for spoken dialog systems. In *Proceedings of Interspeech 2011*.

Michael L. Seltzer, Yun-Cheng Ju, Ivan Tashev, and Alex Acero. 2007. Robust location understanding in spoken dialog systems using intersections. In *Proceedings of Interspeech 2007*, pages 2813–2816.

K. Suzuki and Y. Wakabayashi. 2005. Cultural differences of spatial descriptions in tourist guidebooks. *Spatial Cognition IV. Reasoning, Action, and Interaction*, 3343:147–164.

B. Thomson and S. Young. 2010. Bayesian update of dialogue state: A pomdp framework for spoken dialogue systems. *Computer Speech & Language*, 24(4):562–588.

Ariane Tom and Michel Denis. 2003. Referring to landmark or street information in route directions: What difference does it make? *Spatial Information Theory. Foundations of Geoghraphic Information Science, Lecture Notes in Computer Science*, 2825/2003:362–374.

Barbara Tversky and Paul U. Lee. 1999. Pictorial and verbal tools for conveying routes. In *Proceedings of the International Conference on Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science(COSIT)*. Springer-Verlag London.

Example Dialog

User: Italian restaurant near
ASR: italian restaurant near
NLU: Category=Italian Restaurant

Baseline

Category	*Italian Restaurant*
Target	***Dominos Pizza***

DPOT+Kernels

Category	*Italian Restaurant*
Target	***Dominos Pizza***

User: Italian restaurant near Kappo Nami Nami
ASR: italian restaurant near camp to numa numa
NLU: Category=Italian Restaurant, Street=Camp Avenue
Category=Italian Restaurant, Landmark=Jefunira Camp

Category	*Italian Restaurant*
Street	*Camp Avenue*
Target	***No match***

Category	*Italian Restaurant*
Landmark	*Jefunira Camp*
Target	***Maldonado's***

User: Italian restaurant near Temptations
ASR: italian restaurant near temptations
NLU: Category=Italian Restaurant, Landmark=Temptations

Category	*Italian Restaurant*
Street	*Camp Avenue*
Landmark	*Temptations*
Target	***No match***

Category	*Italian Restaurant*
Landmark	*Jefunira Camp, Temptations*
Target	***Don Giovanni***

Figure 7: Comparison between baseline and proposed method on an example dialog whose underlying true target is an Italian restaurant called Don Giovanni.