Interpretation of Spatial Language in a Map Navigation Task

Michael Levit and Deb Roy

***Abstract*—We have developed components of an automated system that understands and follows navigational instructions. The system has prior knowledge of the geometry and landmarks of specific maps. This knowledge is exploited to infer complex paths through maps based on natural language descriptions. The approach is based on an analysis of verbal commands in terms of elementary semantic units that are composed to generate a probability distribution over possible spatial paths in a map. An integration mechanism based on dynamic programming guides this language-to-path translation process, ensuring that resulting paths satisfy continuity and smoothness criteria. In the current implementation, parsing of text into semantic units is performed manually. Composition and interpretation of semantic units into spatial paths is performed automatically. In the evaluations, we show that the system accurately predicts the speakers' intended meanings for a range of instructions. This paper provides building blocks for a complete system that, when combined with robust parsing technologies, could lead to a fully automatic spatial language interpretation system.**

***Index Terms*—Human–machine interaction, natural language processing, navigational instructions, spatial language understanding.**

I. Introduction

WE PRESENT components of a system that converts verbal descriptions of paths produced by human instruction givers into sequence of actions that an automated agent must take in order to successfully follow paths anticipated by the instruction givers.

Many application areas, including robotics, video games, and geospatial communications analysis, may benefit from an automatic understanding of navigational language. In a video game scenario, for instance, players can be enabled to guide game characters throughout the virtual world of a game. This may be especially powerful when there are large numbers of computer controlled characters, in which case direct control using keyboard and mouse can become cumbersome.

A number of related systems designed to operate in robotic and domestic environment have been described in the literature (e.g., [1]–[5]). In contrast to this previous work that involves sensor-derived (and thus noisy and incomplete) knowledge of the world, we consider the interpretation of relatively complex spatial language by assuming high-level knowledge of the entire map and all landmarks are available to the system.

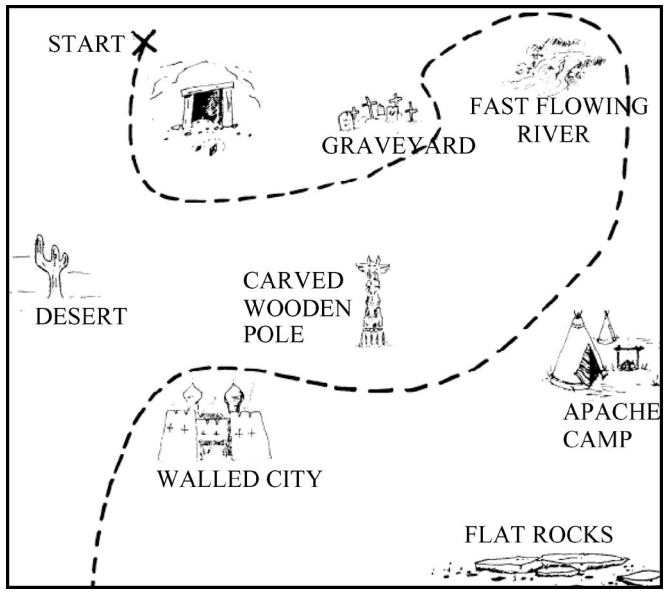

Fig. 1. Sample portion of a map from the MAP-TASK corpus. The path indicated by the broken line only appears on the map seen by the instruction giver. The instruction follower's goal is to recreate this path based on spoken dialog with the instruction giver.

The scenario that we adopted for this paper allows humans to use speech which is unconstrained from both linguistic and representation points of view. The MAP-TASK [6] corpus was selected for system development and evaluation. This corpus is a collection of transcribed human/human dialogs involving cooperative path planning using maps. To collect data, pairs of participants were given similar 2-D maps. One of the participants, the instruction giver, provided navigational instructions to the other participant, the instruction follower, that would guide the latter along a path drawn only on instruction giver's map. An example of a section of such a map with a reference path is depicted in Fig. 1. There were no restrictions whatsoever on language that could be used for navigation. An advantage of this noninvasive "eavesdropping" scenario is that subjects do not attune their navigation strategies to existing or presumed limitations of any automated understanding system (see [7]).

Because of the very high complexity of spontaneous language that arose from the choice of MAP-TASK, we decided to focus on the understanding problem by initially ignoring

Manuscript received February 2, 2006; revised August 15, 2006 and November 19, 2006. This work was supported in part by the National Science Foundation under Grant ITR-6891285. This paper was recommended by Associate Editor E. Santos.

M. Levit was with BBN Technologies, Cambridge, MA 02138 USA. He is now with the International Computer Science Institute, Berkeley, CA 94704 USA.

D. Roy is with the Massachusetts Institute of Technology, Cambridge, MA 02139 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCB.2006.889809

syntactic parsing issues and turning our attention to different basic strategies people used to convey navigational information. Similar to the study in [3], we manually extract basic instructions [which we named navigational information units (NIUs)]; however, our units cover a much broader scope of possible instructions. Some examples of NIUs include moving around objects, moving in absolute directions (e.g., south, left), turning, and verifying closeness to a specific landmark.

One contribution of this paper is in showing that most of these NIUs can be decomposed in a number of "orthogonal" constituents (e.g., type of a move and its reference object), such that the meaning of each NIU or—following a functional approach to understanding—the realization of the path interval it describes can be obtained as a Cartesian product of the meanings of all its constituents. Each of the NIUs can be represented as a parameterized rule with a certain degree of learned flexibility and parameter slots filled by these constituents.

A second contribution of this paper is a novel algorithm that processes sequences of NIUs in order to produce coherent paths which are empirically shown to be similar to the reference paths instruction givers intended to communicate to instruction followers. This integration is possible by virtue of the constraints implicated in the instructions (e.g., moving around an object presupposes that we must be in its vicinity even before the action can take place) and also by some common knowledge (e.g., car–objects cannot be crossed, while bridges can).

II. Navigational Information Units

Extracting basic instruction elements from sentences containing navigational information and grounding them in action primitives are common strategies for understanding systems. The task environment and designer's preferences determine the choice of elements for a particular system, but generally, the idea of splitting instructions in motions and referential descriptions [8] is widely accepted.

In [4], to describe the architecture of a system that understands verbal route instructions in a robotic environment, MacMahon uses four basic instructions: turning at a place, moving from one place to another, verifying view description against an observation, and terminating the current action. While perception undoubtedly plays a crucial role in human orientation and navigation abilities, in our route planning scenario, geometries of all objects participating in a scene are known beforehand, and so we can reformulate the instruction categories above only in terms of this spatial knowledge, thus rendering their procedural aspect more homogeneous.

The feasibility of an automated system that translates from route descriptions to route depictions (and vice versa) is suggested by Tversky and Lee [9]. After studying how humans describe and depict routes, the authors observe that both processes can be decomposed into equivalent sets of verbal and graphic elements, respectively. The lexicon of elements used by the authors consisted of (selecting) landmarks, (changing) orientations, and actions (such as moves), and was borrowed from the study in [10].

In a discussion on the semantics of spatial expressions, Jackendoff [11, Ch. 9] provides linguistic evidence for a conceptual distinction between places and paths. While paths specify trajectories of a traveler, places describe his/its locations. The primary characteristic of a path is the change of location. Turns can be viewed as changes in orientation. These considerations led to four basic types of NIUs in our hierarchy: moves,[1] turns, positions, and orientations. The distinction is not always clear since moving can result in a change of orientation and turning in a practical setting can imply a significant shift in position. We discovered however that, even though different procedures are used to realize moves and turns, the overall path modeling performance does not suffer from the local ambiguity of such issues. Altogether, moves and turns can be subsumed under the general notion of actions, and positions and orientations can be viewed as verifications. Fig. 2 shows the full hierarchy of NIUs. The category compound reference of type auxiliary information is a special type of NIUs that we explain below.

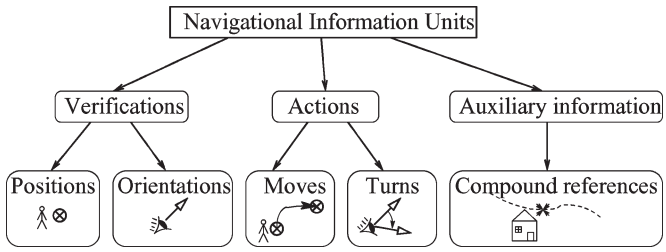

Fig. 2. Hierarchy of NIUs.

Complex spatial instructions are decomposed into a set of NIUs. For example, "now, could you go north past the house till you are right by the forest" is decomposed into the following set of NIUs:

- go north;
- go past the house;
- you are right by the forest.

Currently, human labelers must manually create this decomposition of complex utterances into corresponding NIUs, as well as their constituents (see below). The ultimate goal of this paper is to automate this challenging process of robust parsing and semantic analysis. We do not claim that all of the navigational commands can be classified into the four categories listed above. However, in the experiments, we have found that most of the commands that subjects choose can be classified or decomposed into these categories, and by considering only such commands, we can replicate the paths with reasonable accuracy.

III. Constituents of NIUs

We would like to understand the referential semantics of NIUs extracted from a sequence of sentences that instruction givers say to instruction followers, in order to execute the instructions encoded within. The type of expected system behavior depends on the category of a particular NIU, and for each category, this behavior must be modeled in an appropriate

[1]From now on, we refrain from using the term path in this sense in order to avoid conflict with the notion of path as an end-to-end navigation route.

machine representation. Consider the following move instruction: "move two inches toward the house." Its meaning μ can be decomposed into the following constituents:

$$\begin{aligned}\mu_{\text{move}}(\text{“move two inches toward the house”}) &= \mu_{\text{path descriptor}}(\text{“move ... toward”}) \\ &\times \mu_{\text{reference object}}(\text{“the house”}) \\ &\times \mu_{\text{quantitative description}}(\text{“two inches”}).\end{aligned}$$

If there is a rule for creating a "moving toward" trajectory with respect to a landmark, then we apply this rule to the object which is the meaning of the "the house" expression (its grounding) and follow along this trajectory as far as the meaning of "two inches." In a similar way, we can represent meanings of positions, turns, and orientations.

To reiterate a point made earlier, the currently implemented system processes NIU-constituents, not speech signal, or word transcriptions. While extracting these constituents is a separate research issue that must be addressed in the future (see Section VII), our present goal is to show the viability of this intermediate representation for the understanding task.

In this section, we focus on moves because they represent the most frequent and very informative kinds of instructions, while mentioning other NIU-types whenever a particular constituent is relevant for them. Appendix II illustrates the annotation process by listing all NIUs and their constituents extracted from four consecutive (slightly modified) instruction giver sentences taken from one of the MAP-TASK dialogs.

A. Reference Objects

The first constituent type is a reference object, which denotes an object that serves as an anchor for identifying directions or positions [12], [13]. In our use of move descriptions, the notion of reference objects is broader than just an object with determinable location in space like the ones in [11]. Directions treated as infinitely remote locations encoded in expressions, like "north" or "left," can also be used to describe end points or entire trajectories of moves. The advantage of such an approach will become evident when we consider path descriptor constituents below. There are four major types of reference objects that we have observed in the MAP-TASK corpus: absolute targets, relative targets, landmarks, and compound references.

1) Landmarks are the most familiar class of reference objects; they have finite size and are placed at fixed finite locations. Due to the specifics of the MAP-TASK problem where the objects on the map are drawings on a sheet of paper, we further distinguish the subcategory of page elements referred by expressions such as "page center," "lower edge," "upper left corner," etc., as opposed to genuine landmarks (or simply landmarks): drawings that have prespecified names attached to them (such as FLAT ROCKS or SUSPENSION BRIDGE).
2) Absolute targets are infinite points in space, that are fixed at least for the time of the interaction (for instance, by being tied to the coordinate system of the immobile instruction giver). In MAP-TASK, this is the coordinate system of the map which is oriented in exactly the same way for both instruction giver and instruction follower.[2] Examples of expressions for absolute targets are: "southwest," "down," and "left" (in the sense that is synonymic to "west").
3) Relative targets are infinite spatial deictic references whose meaning changes as the navigation session proceeds. They are used to specify directions from the perspective of the traveler that actually moves along the path and are attached to the traveler's coordinate system. For instance, in the instruction "keep moving," an implicit relative target FORWARD is used, which lies in an infinitely remote point along the traveler's current orientation. Another example is the expression "left," however, this time in the sense of the left side of the traveler's current orientation.
4) Compound references are real or imaginary objects on the map that require an explicit specification in terms of other reference objects; we will deal with them in detail in Section III-E.

Even though the second and third categories of reference objects look more like directions than "objects," they share a very important common aspect with the landmarks: They can be used as anchors to bind move trajectories. Before explaining how this can be done, we note that reference objects are equally important for other NIU-types as well (although not all combinations are possible), e.g., one can "turn to face north" or "be above the house."

B. Path Descriptors

Path descriptors specify how the trajectory of a move is related to its reference object. In [12], Talmy demonstrated that spatial language is schematic insofar as it reduces the information of a scene down to a body of conceptual material assembled on a skeleton of closed-class elements such as prepositions that define spatial relations ("object dispositions") in the scene. Please refer to the study in [12] for a detailed explanation of possible spatial dispositions and how they are constructed using different prepositions. As far as moves are concern, Jackendoff [11] distinguishes four categories: directions with the reference object on a trajectory extension (expressed by prepositions "toward" and "away from"), bounded paths with the reference object in an endpoint of the trajectory (e.g., "from" and "to"), and routes with the reference object related to some interior point of the trajectory (e.g., "via"). We adopt this set of categories, but also extend it to allow each category to be represented by a single rule that we call path descriptor. There are ten path descriptors that are supported by our system (see the upper part of Table I); the trajectory of each of them can be modeled by a circular arc, a straight line interval, or a sequence thereof. For instance, we model a TO-move as a straight line between the current traveler location and the closest point from this

[2]It is certainly true that the real-world orientations of the two maps can be different (instruction giver's west will be instruction follower's east if they face each other), but the crucial fact is that both participants understand each other as long as each of them identifies herself with her map.

TABLE I
PATH AND POSITION DESCRIPTORS. SEE THE MODELING RULES IN APPENDIX I

path/position descriptor	example(s)
TO	*"reach the house"*
FROM	*"leave the forest"*
TOWARD	*"move up"* (abs. ref. obj. NORTH) *"keep going"* (rel. ref. obj. FORWARD)
AWAY_FROM	*"go from east" (to west)*
PAST	*"pass the page center"*
PAST_DIRECTED	*"keep drawing to the left of the rocks"* *"pass right on top of the shack"*
THROUGH	*"follow over the bridge"* *"go across the fields"*
BETWEEN	*"squeeze between the ravine...* *...and the bottom of the page"*
AROUND	*"move around the mill"*
FOLLOW_BOUNDARY	*"follow the lake boundary"*
POS_AT	*"staying close to the beach"*
POS_AT_DIRECTED	*"you are just below the ranch"*
POS_BETWEEN	*"being right between them"*

location that lies on the perimeter of the reference object. In addition, there is one open-end class OTHER to account for all those moves that do not fit in any of the ten classes.

Similarly, it is also useful to introduce position descriptors for position modeling. Currently, our system supports three position descriptors listed with examples in the lower part of Table I.

C. Quantitative Aspect

Some path descriptors, such as TOWARD or AROUND, under specify trajectories in that they encode their shape but do not encode how far the traveler should move. In other words, there is a need for a quantitative aspect in the NIU-descriptions, which would eventually allow a more precise understanding of commands like "move two inches down." With that in place, the traveler will know exactly what to do: select the absolute target SOUTH, extend a TOWARD-move toward it, and follow it for a distance of two inches. Even when a move has an implicit distance specification as in the TO-move "go to the house," the instruction giver may still provide it explicitly ("go one inch to the house"), in which case the instruction follower might need to make some adjustments to accommodate it.

The importance of the quantitative element for direction specifications and problems that arise from it have been addressed by many authors (see for instance the study in [14] and [15]). We distinguish two dimensions in a space of distance specifications. First of all, a distance can relate to a length of the move itself (as in the examples above) or to gaps between trajectories and reference objects (e.g., "pass half an inch above the truck"). Furthermore, there are three distance categories that require different knowledge to the model. Modeling is the simplest when exact units are used: "go about two centimeters to the west." Here, one merely needs to parse the expression "two centimeters" as a measure equal to 2 cm. Such commands are commonly observed in the MAP-TASK corpus. When relative units are used as in "slide down half a page" or "move forward the length of the bridge" (a specification preferred by many authors because it catches relational aspects of distances that define structure of the scene [16]), a more situational competence is required. Finally, in the commands like "keep going for some time" and "move a bit more toward page bottom," the intuitive distance descriptions are used, which demand a significant amount of world knowledge from the interpreter.

Quantitative aspect is also relevant for other NIU types. Therefore, the traveler could be "three inches to the left of the grove" in a position specification or he could "turn forty degrees to the north."

D. Coordinate Systems

Many authors have observed that spatial descriptions are given in terms of a coordinate system in which the scene is taking place. For example, position-NIUs "you are one inch below the house" and "the house is one inch below you" both have reference object "the house" and position descriptor POS_AT_DIRECTED, but their meanings contrast each other clearly because, in the first case, the coordinate system is centered in the house and, in the second case, in the traveler.

From the perspective of cognitive psychology, the most important question about coordinate systems is whether the system is bound to the experiencer (egocentric[3] coordinate system) or is independent of her (allocentric coordinate system) [17]. These two major categories can be further subdivided according to where exactly the coordinate system is centered and what it uses as a reference object. Regarding spatial deictic references, Levelt [13, Ch. 2] distinguishes among the following three major cases.

1) Primary deictic reference: Here, the speaker is the origin of the coordinate system and also the reference object (relatum) (example: "the ball is in front of me").
2) Secondary deictic reference: The speaker is the origin of the coordinate system, but not the reference object: "the ball is behind the tree."
3) Intrinsic reference: The reference object (not speaker) is also the origin of the coordinate system; here, the reference object must possess its own "intrinsic" orientation with front and back: "the ball is in front of the house" (see also [12, p. 241]).

Similar categorization suggestions can also be found in [18] and others. By virtue of the examples above, we could see that orientation is indeed important when defining a coordinate system. As an arbitrary coordinate system is defined by: 1) its origin; and 2) its orientation, our approach to the spatial language in MAP-TASK is to organize all possible coordinate systems into a 2-D grid presented in Table II. Here, there are two possible origin placements and three different orientation types for the coordinate systems in which NIUs can be specified.

There are three possible perspectives in the MAP-TASK: one of the instruction giver, one of the map traveler (often identified with the instructions follower), and finally, a perspective from some reference object on the map (landmark or page element). However, only two of them can have an origin associated with

[3]In reality, the egocentric system itself is hypothesized to be an acquired complex coordination of several sensory-motor manifolds [17].

TABLE II
TYPES OF COORDINATE SYSTEMS IN WHICH NIUS CAN BE SPECIFIED WITH NIU-EXAMPLES

		orientation		
		absolute	traveler	reference object
origin	ref. object	*"pass to the left of the page center"* (move)	*"descent along your side of the rocks"* (move)	*"you are at the base of the monument"* (position)
		"turn to face north" (turn)	*"go to the other side of the lake"* (move)	*"slide a few inches down the river"* (move)
	traveler	*"go due southwest"* (move)	*"this house should be on your right"* (orientation)	?
		"the lake above you" (position)	*"turn around"* (turn)	

them since the instruction giver is not really located "on the map" and can only define orientation. There exists a certain redundancy in the choice of a coordinate system and specification of reference objects. For instance, whenever the reference object of a move is a relative target, the coordinate system is always placed where the traveler is and oriented according to the traveler's orientation. Besides, the exact specification of a coordinate system or at least of its orientation can be irrelevant for certain NIU-types. For instance, in the position-NIU "it is close to the barn," orientation of the coordinate system (which is placed in "the barn") cannot be determined and is, in fact, not needed for understanding.

E. Compound References

If the language of instruction givers were constrained to the kind of examples we have seen before, it would have only a limited expressive capacity because it would not allow for a very large portion of potential reference objects to be taken into account. The set of landmarks and page elements is too sparse and lacks the needed expressive means to allow for high-precision navigation. Instructions such as "go to the lake" under specify the required action because the lake can occupy a large portion of a map. Instead, something like "go to the northwest corner of the lake" is needed. Similarly, if the target of a particular TO-move is a point an inch above the house, there is no way to avoid an explicit specification of this point, "move to a spot slightly above the house," or even simpler, "move above the house." All these are examples for what we call compound references. Nested compound references are also possible: "continue toward [the spot an inch under (the bottom of the monument)]" as well as compound references that have stretch ("you are level with the springs").

The description language for compound references is very similar to the one of positions, except for one special case where the compound reference is a part of its own reference object as in "you should be right under the gate of the castle," and so (even though their semantic role is clearly different from that of actions and verifications) we decided to include compound references in the list of supported NIU types (see Fig. 2). In the present version, we are not trying to estimate positions of compound references; instead, we assume they are known and thus look them up in manual annotations.

F. Designing and Validating Rules

Up to now, we have largely ignored the question of how to model individual NIUs, concentrating mainly on the possibility of such modeling. This section describes how we design and use a lexicon of action and verification primitives.

The first part of the process is a manual step of designing prototypes for each rule. Then for each NIU type, we compare the designed prototype with actually observed instances[4] in order to estimate spatial templates [19]. Later, when path intervals corresponding to individual NIUs are merged together to form a continuous replica of the reference path, these spatial templates will guide this process, which helps in finding the most probable realization for each NIU that allows for such a merge.

We first consider position-NIUs of type POS_AT_DIRECTED. There is an extensive prior work on modeling spatial language (e.g., expressions "above," "to the left of," etc.) [19]–[21]. We chose an easily implementable model similar to the hybrid model of the study in [21] to model these NIUs. Two metrics determine goodness of a particular position with respect to its reference object: first, the angle between the defining axis (e.g., vertical axis in case of "above") and the beam emanating from the reference object's center of mass and passing through the position and, second, the projection of the distance from the extreme point of the reference object in the given direction (e.g., the highest point for "above") to the position on the defining axis (e.g., distance of y coordinates for "above"). For NIUs of type POS_AT, there is only one metric: absolute distance from the position to the closest point of the reference object.

For moves and turns, creating spatial templates is similar: First, for each actually observed move/turn, we seed its corresponding prototype into its starting point and, thus, obtain its predicted version. Second, the radial and angular deviations between the end points of the predicted and observed path intervals are computed. Prototypes and intuitive explanations of these distances for moves of types TO and FOLLOW_BOUNDARY are shown in Fig. 3. Here, we execute a FOLLOW_BOUNDARY-move by "expanding" the perimeter of the reference point to traverse the current traveler position and moving along this expanded perimeter in the direction (of the two possible) that has the smallest angle with the orientation that the traveler had before reaching her current position.[5] For a detailed investigation of when and how people use path descriptors of this type as opposed to path descriptors of type PAST, please refer to the study in [22].

In short, radial distance is the difference in lengths of the observed and predicted moves, and the angular deviation shows how far from the predicted trajectory the actual trajectory deviates. For some of the prototypes (like, for instance, TOWARD-moves), we need to provide not only the path descriptor but also the default distance. In our experiments, this distance was estimated empirically.

[4]As we have mentioned earlier, these instances (path intervals) are part of manual annotations that we create prior to the experiments.

[5]In practice, we used a computation scheme for angular and radial deviations for FOLLOW_BOUNDARY, which is slightly different from the depicted one and approximates it instead.

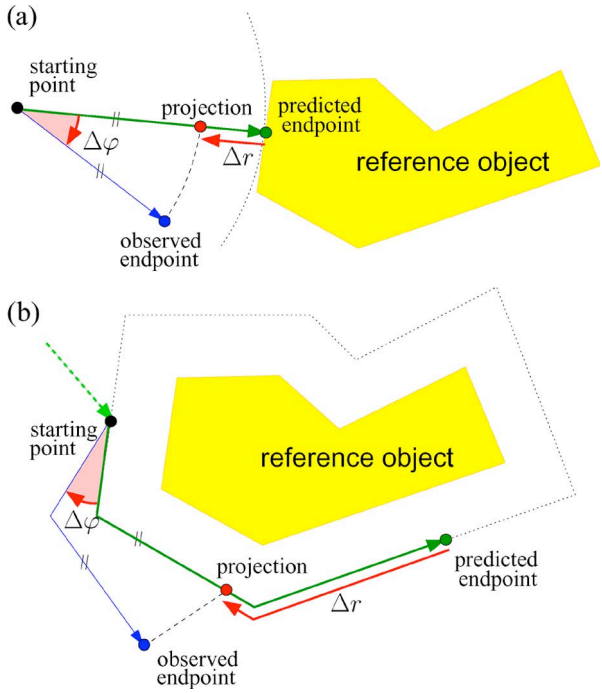

Fig. 3. Prototypes and radial and angular deviations for moves of types (a) TO and (b) FOLLOW_BOUNDARY.

Angular and radial deviations collected for each NIU type are compiled into a 2-D spatial template. This template contains probabilities of all realizations of this NIU type that start in the same point but deviate from the prototype in due course. This is also why there is no need for searching for the perfect prototype for each NIU type. Indeed, the spatial template will compensate for potential mistakes.

To summarize, the models of individual NIUs are combinations of handcrafted structures combined with data-driven parameter adaptation that make these models flexible. Positions are 2-D probability distributions for all locations, and moves and turns are 2-D probability distributions as well, but for the end points of actions they represent (and indirectly also for different trajectories that lead to these endpoints). In experiments reported below, we ignore orientation-NIUs because they occur fairly seldom in the corpus.

IV. Combining NIUs Into Contiguous Paths

Now that we have described how NIUs are grounded in particular action and verification primitives, we turn to the issue of combining sequences thereof into a contiguous path on the map. This task can be considered as the one of route planning, where instructions are given before the actual following of the route takes place.

A. Dynamic Programming Approach

The "plan-as-communication" view on plans in [23] suggests that plans constrain possible space of actions and require some interpretative effort from the agent whenever execution of a particular action is due; in other words, it enforces by considering each action in a context of the whole plan and of the environment the plan operates in. Our approach is motivated along similar lines. In our system, interpretation of instructions like "turn left" or "go to the house" takes place only when they are next to be executed, and probabilistic assessment of geometries of their reference objects with respect to the traveler's current position and orientation can be made. Only then "left" and "go to" will, for the first time, acquire a concrete meaning attached to them. However, this meaning is by no means final, for instructions that follow can still change it later on. If we do have a map in front of us at the planning time, we can mentally follow the route right away and "rehearse" execution of all navigational instructions in the sequence one by one. If at some point in time we realize that a mistake has been made on previous stages because no consistent continuation of the path is possible, we can always back-off to the point where the mistake was made and choose other alternatives leading from there. Moreover, we can simultaneously maintain several alternative routes in the first place, scoring and extending them in parallel as we progress and dynamically preferring one of them over the others. This view of the task suggests a dynamic programming approach.

Dynamic programming, however, cannot operate on a continuum of $\mathbb{R}^2$, which is the case for maps in MAP-TASK, but rather needs a set of discrete alternative states. In order to achieve that, we impose a rectangular grid on the maps and consider only cell centers as potential alternatives for traveler's locations at the end of each action.

At this point, let us restrict the NIUs to moves and turns only and assume that the entire path is split in N intervals; each of which is covered by exactly one NIU.[6] Let us also assume that our map is split in I square cells c_i, $i \in \overline{1, I}$. Then, on each step $n \in \overline{0, N}$, there will be a separate probability distribution ending up in cell $c_i \forall i \in \overline{1, I}$ after this step has been taken. Since the starting point of the path is considered given, the initial probability distribution $(n = 0)$ is 0.0 for all cells except the one containing the starting point, where it is 1.0.

Assume now that we know the probability distribution $p_n(j)$ on step n. Conditional probabilities $p_{\gamma_n}(i|j)$ of ending NIU γ_n in c_i, given that it starts in c_j, can be interpolated for all $j \in \overline{1, I}$ from spatial templates that have been discussed in the previous section (see Fig. 4).

Then, the total probability of reaching c_i on step n and passing through c_j before that is

$$p_{n+1}(i, j) = p_n(j) \cdot p_{\gamma_n}(i|j). \tag{1}$$

With the decision-oriented approach to probabilities, we select the predecessor index j^*

$$j^* = \arg\max_j p_{n+1}(i, j) \tag{2}$$

such that cell c_{j^*} is the predecessor of c_i on the optimal path and declare

$$p_{n+1}(i) := p_{n+1}(i, j^*). \tag{3}$$

[6]We will show later how this unrealistic assumption can be removed.

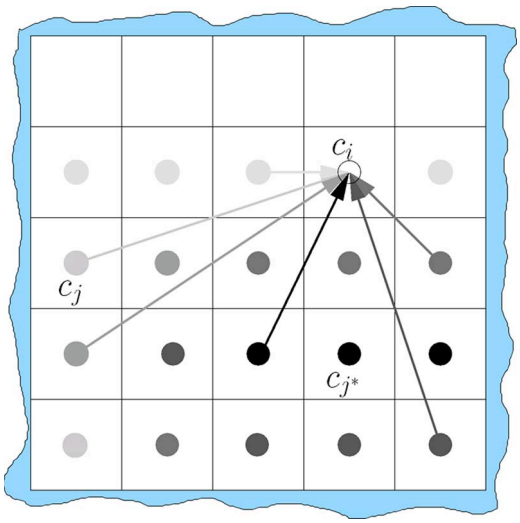

Fig. 4. Computing probabilities of endpoints for NIU: "go north." For each cell c_j, the probability of starting there and ending up in cell c_i is encoded as the darkness of the corresponding arrow and of the cell center. It depends on the deviation from the north direction as well as on the distance between c_j and c_i.

After all N NIUs have been processed, the entire optimal path must be recovered. For that, we start in the cell c_{j^*} from the last distribution and traverse backward the sequence of the distributions for all NIUs, following the line of predecessors all the way up to the very first distribution.

It is important to see that, in this simplified task formulation, the "winning" cells describe not only the most probable position after having processed the last NIU but also the orientation of the traveler that goes along with this particular NIU realization.

B. Natural Task Constraints

At first sight, it might appear unclear why spatial templates are needed at all. Indeed, if no constraints existed in the task, for all n's, the nth step of the dynamic programming algorithm would always result in selecting the realization of γ_n that possesses angular and radial deviations corresponding to the maximum in γ_n's spatial template. Fortunately, there are, in fact, several constraints that come along with the choice of the task domain that we call "natural" constraints of the task. In general, we can say that these constraints are task dependent and arise from the working conditions of the system.

First, our domain expertise and intuition suggest that some of the landmarks cannot be crossed. The maps designed for MAP-TASK comply with this consideration to a great extent. For example, while a path can pass through the drawing of a bridge, it will never cross a rock. This knowledge is one of the main sources of constraints that shape possible paths. In the same way, we might want to prohibit self-crossings of a path and restrict all the paths to the inside of the visible map. One should keep in mind, however, that banning self-crossings hurts the optimality principle of the dynamic programming saying that solutions of partial problems never need to be recalculated [24] and thus can lead to not finding a good path.

Second, we propose that a reasonable bias is to favor junctions that are smooth, i.e., large changes in orientations when finishing modeling γ_{n-1} and starting γ_n should be penalized. In particular, this bias will eliminate abrupt near-180° turns.

Finally, for many move types that have landmarks for reference objects, it is reasonable to presume spatial proximity of the starting point to the landmark. In fact, in the MAP-TASK corpus, there are, rarely, instructions to move around some landmark that is on the other end of the map, far away from our current position. This means that only those realizations of such an NIU that ended close to the landmark will be considered in the next step. Indirectly, these constraints are modeled in spatial templates; however, we found out that imposing explicit upper thresholds on maximum distances between starting point of an NIU and its reference point is helpful as well. Besides, we can require a certain degree of consistency from action trajectories and end points. For instance, the end point of a BETWEEN-move must, indeed, lie between its reference objects, and the PAST_DIRECTED-move expects the instruction follower to be on a particular side of the reference object.

C. Integrating Positions

Another source of constraints is verification-NIUs, in particular positions. They too restrict the working space to a small area tied to a reference object or (in the case of POS_AT_DIRECTED) to one of its sides. Consider how positions can be integrated in the framework we have developed so far. Recall that we update the distribution of locations after each action. Similarly, we can update them after each position specification as well. Here, however, we can multiply the position probabilities $p_{\gamma_n}(i)$ (interpolated from the corresponding spatial template) with the distribution $p_n(i)$ obtaining a new adjusted distribution $p_{n+1}(i)$ as

$$p_{n+1}(i) : p_n(i) \cdot p_{\gamma_n}(i). \tag{4}$$

Even though we do not model rare orientation-NIUs in the experiments of this paper, they can be handled in exactly the same way as positions.

D. Dealing With Redundancy

No matter how many position specifications there are, all of them can be subsequently treated as shown in (4). This, however, is not true for actions. If several action-NIUs compete for one path interval or even describe path intervals that only start in approximately the same location, their contributions must be considered simultaneously. Let $\Gamma_n = \{\gamma_n^k\}, k \in \overline{1, K_n}$ be a set of NIUs competing to define the next path interval. In order to compute joint probabilities $p_{n+1}(i, j)$, we average over individual NIUs in Γ_n and, assuming equal priors for all NIUs in the set, modify (1) into

$$p_{n+1}(i, j) = p_n(j) \frac{1}{K_n} \sum_k p_{\gamma_n^k}(i|j). \tag{5}$$

After that, selection of the optimal predecessor and computation of the next distribution $p_{n+1}(i)$ is done as before.

In contrast to the remark at the end of Section IV-A, the resulting orientation after Γ_n is yet to be determined since it

is a product of several NIUs at the same time. Our approach to this problem is to select the NIU that delivers the highest conditional probability $p_{\gamma_n^k}(i|j^*)$ to represent the group, but to use the orientation computed as a weighted average over all NIUs in Γ_n to control smoothness of the path.

One issue we have not addressed yet is how to establish a partial relation on all NIUs of a session, i.e., which NIU should be considered when, and what are the sets Γ_n of competing NIUs. For now, our system looks up this information in the annotations, looking at the starting points of all NIUs.[7] This information is usually contained in the language, and from proximity of NIUs in dialog transcriptions, we can usually conclude at least on proximity of the intervals they describe. For instance, the sentence "go left to the creek" contains two NIUs, "go left" and "go to the creek," that should be put in the same set Γ_n. A more sophisticated linguistic analysis is required if precedence must be determined as well (see Section VII). In other situations where the instruction giver comes back to one of the already described path intervals reiterating or rephrasing extractions given earlier, there are dialog context clues (e.g., instruction follower's feedback) to signal this fact.

We employed one general rule regarding splitting the set of NIUs with close starting points, that resulted in performance improvement. If there are verification-NIUs, we first create a set out of them. Then, if there are action-NIUs that are reliable in guessing directions, such as TO- and TOWARD-moves, and turns with absolute targets or landmarks as reference objects, we make a separate set out of them, and process this set only after the first one. Next, a set of other moves with such reference objects is processed. Finally, if and only if no action groups could be created, actions with relative targets as reference objects are considered.

V. Evaluation Metrics

There are two classes of evaluation metrics that are of interest for this paper. The first class of instruction-level metrics concerns modeling of individual NIUs and sheds light on the quality of path descriptor rules by assessing deviations of observed actions around their prototypes. The second class of path-level metrics evaluates the entire paths and judges their overall quality by comparing them to their references on the instruction givers' maps.

On the instruction level, we can judge shapes and, in particular, compactness of spatial templates (distributions of angular and radial deviations). Visual assessment is important for the entire paths; however, we can also use criteria such as percentage of landmarks on a correct side of the path and average trajectory deviations to perform their formal evaluation. A reasonable figure of merit for the latter is the area between the observed and predicted paths. If we augment the predicted path so that it ends in exactly the same point where the reference path ends (human objects in the original MAP-TASK experiment knew the position of the finish), then the two paths together will form a closed contour, and we can use standard filling techniques such as SCAN LINE algorithm [25] to compute the cumulative area of the "mismatch"-zones (see Fig. 5). The smaller the area is, the better is the model.

[7]Note that we do not look up the exact positions of these starting points on the path, but rather only the fact that they are close for two or several NIUs.

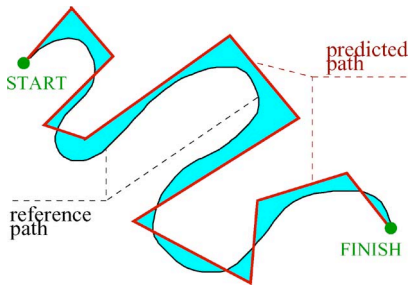

Fig. 5. Area between the observed reference path and predicted paths can be used to assess quality of modeling.

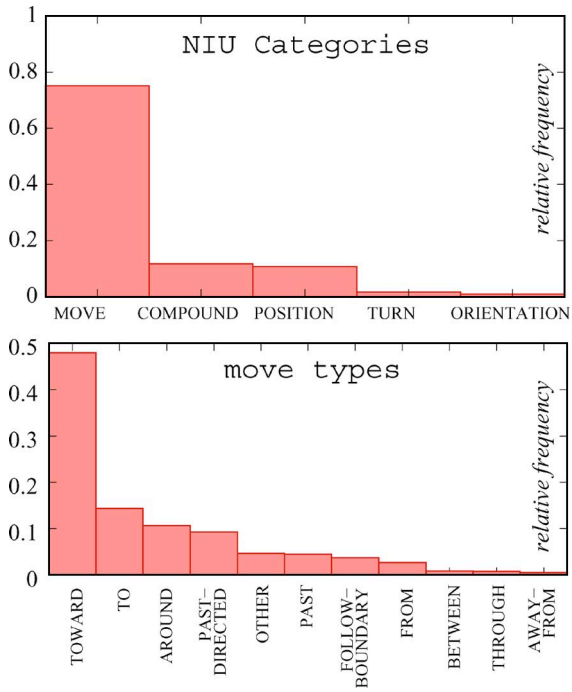

Fig. 6. Occurrence statistics of NIU categories and move types.

VI. Results

Our experiments were conducted on the commercially available HCRC MAP-TASK corpus [6]. This corpus consists of 128 navigation sessions with audio recordings and a number of different annotations available in XML-format for each session. We randomly selected 25 of these sessions for our experiments and focused our analysis on instruction givers' speech. Based on a set of previously existing annotations of this speech data in terms of moves in conversational games [26], we selected the subset annotated as either instructions or clarifications. These sentences were then manually annotated with respect to the NIUs they contain. We defined the NIUs reported in this paper on the basis of analyzing only five of the 25 sessions. On average, we obtained 85 NIUs per session. Relative frequency distributions of categories of the extracted NIUs as well as of move types within the move category are shown in Fig. 6. These

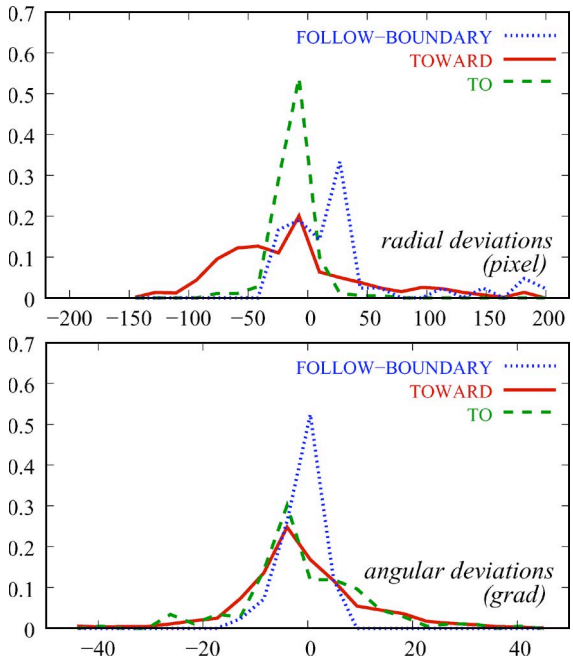

Fig. 7. Angular and radial deviation statistics estimated for moves of types FOLLOW_BOUNDARY, TOWARD, and TO.

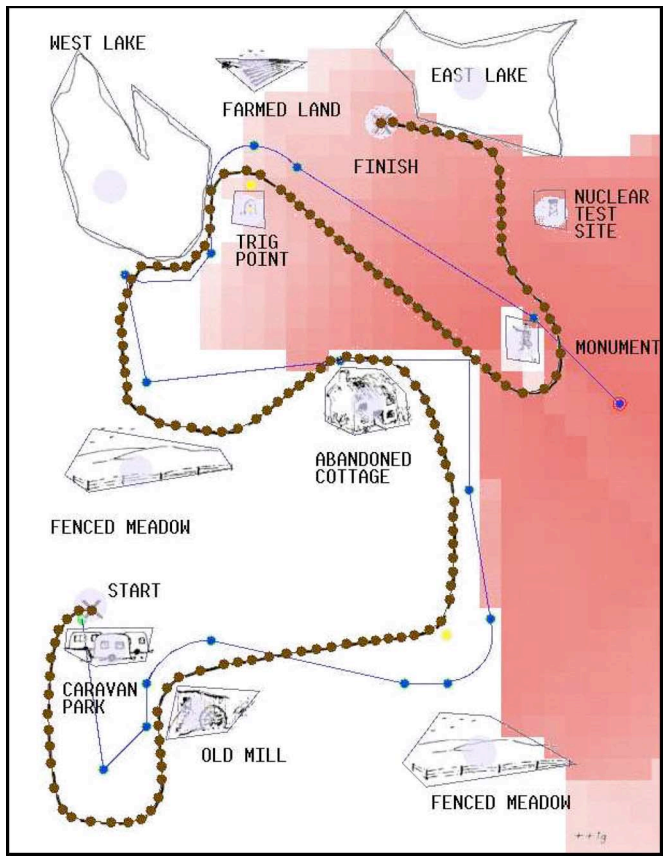

Fig. 8. Snapshot of a modeled path after processing several groups of NIUs.

plots show that moves clearly dominate among all analyzed NIUs and the TOWARD type, that includes instructions such as "draw your line toward the northeast," "move down," "keep going," etc., is the most frequent type among moves. Less than 5% of the 2133 annotated NIUs could not be identified as one of the five NIU categories from Fig. 2, and less than 5% of the 1526 annotated moves have been labeled with the path descriptor OTHER. The high coverage of these NIUs for the complete set of 25 sessions suggests that this set of NIU-models is well suited to the task and perhaps also spatial language for navigation tasks, more generally.

Another reassuring confirmation comes from the following measurements. Interannotator agreement with respect to the extraction of NIUs and their labeling with one of the five supported types as well as with respect to path descriptors of detected moves was roughly estimated for two labelers using the F-measure. It amounted to 0.86 and 0.8, respectively, with the absolute majority of mismatches due to ambiguities of cases like "go above the house," which can be interpreted either as a TO-move with a compound reference object or as a PAST_DIRECTED-move. Nonetheless, our experiments showed that such ambiguities do not impair the understanding of complete paths.

We then produced discrete versions of the reference paths, representing them as sequences of many "stops" placed densely along the original curve. Each NIU was annotated with a path interval (delimited by the first and last stops on the path) that it, in labelers' view, accounts to. Based on these annotations, we estimated spatial templates for each of the move types, position types, and turns, expressing them in terms of radial and angular deviations from manually designed prototype rules. As expected, the main source of deviations came from the underdetermined quantitative aspect of NIUs; for example, in Fig. 7, we see that, while estimated angular deviations statistics possess rather compact distributions (meaning that the rules we designed to represent these move types are in fact consistent with annotations), radial deviations of the moves (variations of their stretches) along the given trajectory are flatter for those move types that intuitively require explicit stretch specification (such as TOWARD-moves).

Next, we show how the dynamic programming approach can be used to integrate models of individual NIUs in a joined consistent and smooth path. A snapshot of one of the replicated paths (in progress) is shown in Fig. 8. Several aspects of this path are noteworthy. First, the snapshot sheds light on the way the drawing is discretized. For each landmark, we marked its perimeter that cannot be crossed unless there is no other way to proceed with the path. This otherwise exceptional case happened to occur here at the beginning, where the instruction giver insisted on going down and the spatial template for a TOWARD-move prohibits angular deviations of more than 50°. Also, the rectangular grid of cells for which a new distribution is estimated after each processed group of NIUs can be seen here: The darker the cell is, the higher is the probability of ending up there; the cell with the highest probability is chosen to determine the most probable path so far (sequence of circles and lines and arcs between them). The distribution in this snapshot takes place after one NIU that sends the traveler a specified distance toward the southwest and another one that commands to go on along the same direction. From this distribution, it can be seen that self-crossings are prohibited and perplexity of

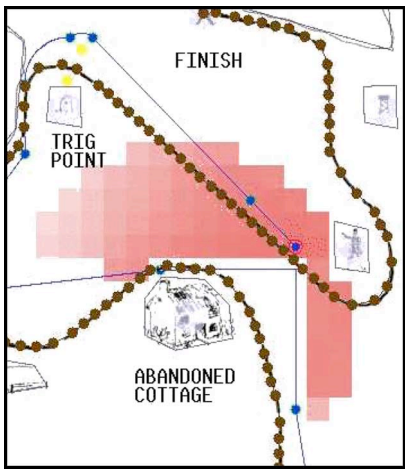

Fig. 9. Excerpt of the same path after the next instruction (position-NIU "near to the abandoned cottage").

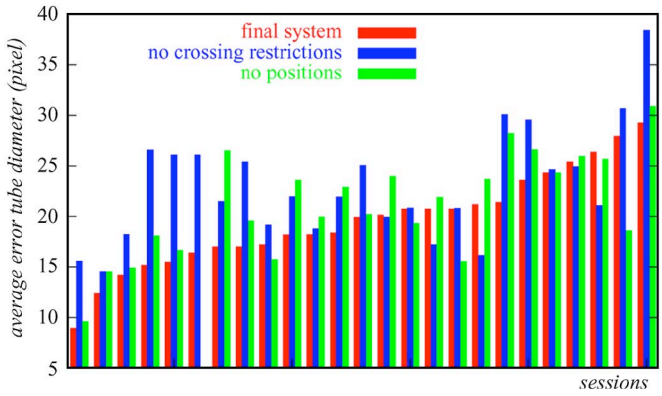

Fig. 10. Error tube diameters for all sessions computed for cases where all available information is used, no landmark and self-crossing restrictions are imposed, and only move- and turn-NIUs are considered.

such a distribution can get very high. One of the possibilities to reduce the perplexity is to issue a verification-NIU. In the presented session, the next instruction was indeed position-NIU "near to the abandoned cottage," and the new distribution with a lower perplexity resulting from it is shown in the excerpt of the map in Fig. 9. It is more compact, with the only allowed cells located around the landmark.

As far as the quality of the predicted path, it can be seen that it lies reasonably close to its reference path. In order to formalize the visual assessment, we computed areas of the "mismatch"-zones for each pair of reference/predicted paths (see Section V) and normalized them by the length of the corresponding reference paths. This criterion can be interpreted as the average diameter of an "error tube" of deviations that we enwrap the reference paths into. The smaller the error tube diameter is, the more precise is the modeling mechanism. Evaluation of the predicted paths for all 25 sessions resulted in an average diameter of 19.5 pixels (one fortieth of the height of the maps) with a sample deviation of five pixels. For comparison, the baseline strategy of connecting start and finish landmarks with a straight line resulted in an average error tube diameter of 280 pixels. We take this as a clear sign of success for our modeling algorithm. Unfortunately, the original corpus contained no paths drawn by instruction followers on their maps. This kind of information would have provided another important baseline for our experiments.

In order to investigate the importance of the natural constraints and verifications for a successful path modeling (Section IV-A), we conducted one replicating experiment without any restrictions on landmark and self-crossing, and another one where all the position-NIUs were ignored.[8] For the first experimental setup, we obtained an average diameter of the error tube of 23 pixels (sample deviation is 5.5 pixels). For the second one, one session could not be completed at all, and for those that could be completed, we obtained an average diameter of 21 pixels (sample deviation is five pixels), which amounts to a relative precision loss of 18% and 7%, respectively. In terms of a number of landmarks passed on the wrong side, removing all position-NIUs increased their proportion by almost 50% relative. All of the above experiments were conducted using leave-one-out strategy, i.e., in order to replicate each session, we trained the spatial templates on the remaining 24.

These results demonstrate the importance of natural constraints and verifications in navigational tasks. Fig. 10 shows the error tube diameters for all sessions for all these experiments where sessions are arranged in such an order that the diameter increases for our final system with no landmark and self-crossings allowed and with position-NIUs accounted for.

Another promising result comes from ignoring stretch specifications for moves. In the previous experiments, if the quantitative constituent of a move was specified, we would temporarily shift a corresponding spatial template to peak in this stretch. However, looking up the meaning of expressions like "a little bit" is not quite fair because it requires serious semantic analysis and a great deal of world knowledge. As it turns out, we can ignore such explicit specifications altogether, and the integration procedure will still deliver accurate models. In our experiments, the average error tube diameter remained under 20 pixels.

VII. Discussion

We have reported first steps toward automatic understanding of unconstrained navigational instructions in the MAP-TASK domain. Clearly, substantial aspects of the problem remain unmodeled and pose significant challenges for future research. For example, we still need to extract NIUs from instruction transcripts, impose a partial precedence relation on them, and understand the meaning of distances and angles depending on manual interpretation as well. As far as the latter is concern, we showed in the previous section that the quantitative aspect of NIUs can be ignored without significant loss in performance if we consider them in context of other NIUs. Extracting NIUs from text is a task similar in spirit to the task of named entity extraction and may be achieved using well-established tagging algorithms [27]. Our preliminary experiments in this direction

[8]The remaining moves still had a number of constraining elements (like PAST_DIRECTED- or TO-moves), so that the modeling did not break apart.

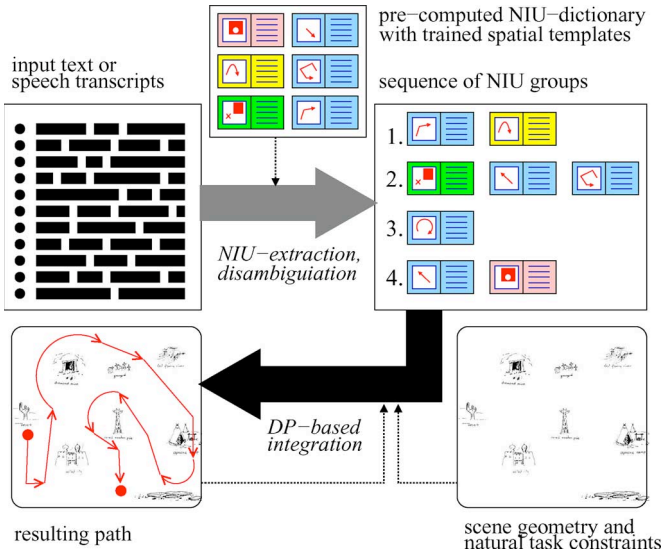

Fig. 11. Automatic path replication. System diagram.

produced promising results (not reported in this paper). Determining the partial order of NIUs would remain a challenge. For instance, the instructions "at the corner go left" and "go left till you are at the corner" both contain one position describing being at the corner and one move describing going to the left. But in the first case, the position precedes the move, and in the second, it is the inverse. Ultimately, more robust syntactic processing must be brought to bear on the problem. We believe that our approach of propagating spatial constraints based on NIUs in a dynamic programming framework provides an extensible framework for such future investigations.

To put the work completed thus far into context from a practical perspective, we sketch how the implemented components might play a role in a larger end-to-end language understanding system. Extraction of NIUs from text or speech transcripts is the only transition that has not been completely automated yet; it is identified by the thick gray arrow in Fig. 11. Given a sequence of extracted groups of competing NIUs whose spatial templates are learned from a training corpus, we then use the DP-based algorithm from Section IV-A to produce a contiguous path, while also taking into account scene geometry and the scenario's natural constraints. In a video game scenario mentioned in the beginning of this paper, for instance, our system can reside inside an action interpreter that reads multistep natural language path descriptions submitted by the players via keyboard or microphone and sends virtual game characters to follow the complex trajectories that arise from them.

VIII. Conclusion

We have described a system that infers paths on maps by processing natural language instructions represented as NIUs. This translation process from linguistically derived symbolic representations to geometric spatial representations is an example of language grounding (see [28] and [29]). Our focus in this effort was to automate the translation of NIUs into probability distributions over possible paths on maps. We defined four categories of NIUs, moves, turns, positions, and orientations, and developed an approach for composing NIUs in order to interpret the semantics of complex natural utterances that are analyzed as comprising multiple NIUs. In the evaluations, this approach successfully produced semantically correct interpretations for a wide range of utterances.

Appendix I
Definitions of Path and Position Descriptors

This section explains the path and position descriptors that can be used to classify moves and positions, as well as turns.

1) TO: in a straight line, approach the closest point of a reference object.
2) FROM: by keeping the previous direction, make sure the move goes away from a reference object.
3) TOWARD: move in the direction of the center mass of a reference object.
4) AWAY_FROM: move in the direction opposite to center mass of a reference object.
5) PAST: by keeping the previous direction, proceed in a straight line up to the point where the farthest point of a reference object projects on this direction.
6) THROUGH: in a straight line, proceed through the center mass of a reference object and up to its farthest point in this direction.
7) PAST_DIRECTED: this path descriptor can have one of the following four subcategories (sides): "above," "below," "to the left of," and "to the right of" a reference object. It consists of one or two straight line intervals. If the traveler is not already on the required side of a reference object, he has to take the shortest path to get there (possible directions: north, south, west, and east). The second step leads from the past projection of the center mass of a reference object on the required side to a projection of the farthest point of the reference object on it.
8) AROUND: in a circular arc, move around the center mass of a reference object; among two possible initial directions, select the one closest to the previous direction.
9) FOLLOW_BOUNDARY: "expand" the perimeter of a reference object to pass through the starting point of the move. Follow this expanded perimeter; among two possible initial directions, select the one closest to the previous direction.
10) BETWEEN: this move requires two reference objects. Compute intervals of view angles not crossing any of the reference objects and consider two directions in their middles. Select the one closest to the previous direction and proceed in a straight line up to the projection of the farthest point of both reference objects on this direction.
11) TURN: turns are modeled similar to the AROUND-moves with a small radius arc. Traveler follows the arc till needed orientation is achieved.
12) POS_AT: this position descriptor generates score that depends on the traveler's distance to the closest point of a reference object.
13) POS_AT_DIRECTED: similar to PAST_DIRECTED, there are four possible sides for this position descriptor

("above," "below," "to the left of," and "to the right of"), each one determining its active direction (e.g., north for "above"). The score depends on the traveler's distance along the active direction to the reference point's extreme in this direction and also on the angle that a beam from the reference point's center mass to the traveler creates with the active direction.

14) POS_BETWEEN: here, the score is generated based on a difference between distances from the traveler to the closest points of the first and second reference objects.

APPENDIX II
ANNOTATION EXAMPLES

Consider the following four (slightly modified) instructions from one of the MAP-TASK dialogs (see also Fig. 8).

- "Continue up north slightly."
- "... to the tip of the lake."
- "... and then we're going to turn down above the trig point."
- "... and we're going to turn immediately to your right."

For these sentences, the following NIUs have been annotated.

1) MOVE with path descriptor TOWARD ("continue"), absolute reference object NORTH ("up north"), and intuitive stretch from start ("slightly").
2) COMPOUND REFERENCE of the type PART-OF ("the tip of").
3) MOVE with path descriptor TO ("to") and the above compound reference as a reference object.
4) TURN with absolute reference object SOUTHEAST ("down," southeast direction observed on the map).
5) POSITION with position descriptor POS_AT_DIRECTED ("above") and coordinate system with a center in the reference object TRIGPOINT ("trig point") and absolute orientation.
6) TURN with relative reference object RIGHT ("your right").

ACKNOWLEDGMENT

The authors would like to thank all the members of the Cognitive Machines group for sharing their views on the task and providing useful remarks regarding the methods and algorithms employed in this research, and S. Tellex, R. Kubat, and the anonymous reviewers for their comments on the earlier drafts of this paper.

REFERENCES

[1] R. Simmons, D. Goldberg, A. Goode, M. Montemerlo, N. Roy, B. Sellner, C. Urmson, A. Schultz, M. Abramson, W. Adams, A. Atrash, M. Bugajska, M. Coblenz, M. MacMahon, D. Perzanowski, I. Horswill, R. Zubek, D. Kortenkamp, B. Wolfe, T. Milam, and B. Maxwell, "GRACE: An autonomous robot for the AAAI robot challenge," *AI Mag.*, vol. 24, no. 2, pp. 51–72, 2003.

[2] C. Miranda-Palma and O. Mayora-Ibarra, "Robotic remote navigation by speech commands with automatic obstacles detection," in *Proc. Robot. and Appl.*, 2003, pp. 53–57.

[3] G. Bugmann, E. Klein, S. Lauria, and T. Kyriacou, "Corpus-based robotics: A route instruction example," in *Proc. IAS-8*, Amsterdam, The Netherlands, Mar. 2004, pp. 96–103.

[4] M. MacMahon, "MARCO: A modular architecture for following route instructions," in *Proc. AAAI-05 Workshop Modul. Constr. Human-Like Intell.*, Pittsburg, PA, Jul. 2005, pp. 48–55.

[5] S. Tellex and D. Roy, "Spatial routines for a speech controlled wheelchair," in *Proc. HRI*, 2006, pp. 156–163.

[6] A. H. Anderson, M. Bader, E. G. Bard, E. H. Boyle, G. M. Doherty, S. C. Garrod, S. D. Isard, J. C. Kowtko, J. M. McAllister, J. Miller, C. F. Sotillo, H. S. Thompson, and R. Weinert, "The HCRC map task corpus," *Lang. Speech*, vol. 34, no. 4, pp. 351–366, 1991.

[7] T. Tenbrink, K. Fischer, and R. Moratz, "Spatial strategies in human–robot communication," in *Themenheft Spatial Cognition*, C. Freksa, Ed. Bremen, Germany: arenDTaP Verlag, 2002. KI 4/02.

[8] C. Riesbeck, "You can't miss it: Judging the clarity of directions," *Cogn. Sci.*, vol. 4, no. 3, pp. 285–303, 1980.

[9] B. Tversky and P. U. Lee, "Pictorial and verbal tools for conveying routes," in *Spatial Information Theory: Cognitive and Computational Foundations of Geographic Information Science*, C. Freksa and D. M. Mark, Eds. Berlin, Germany: Springer-Verlag, 1999, pp. 51–64.

[10] M. Denis, "The description of routes: A cognitive approach to the production of spatial discourse," *Curr. Psychol. Cogn.*, vol. 16, no. 4, pp. 409–458, 1997.

[11] R. Jackendoff, *Semantic and Cognition*. Cambridge, MA: MIT Press, 1983.

[12] L. Talmy, "How language structures space," in *Spatial Orientation: Theory, Research and Application*, H. L. Pick, Ed. New York: Plenum, 1983.

[13] W. J. M. Levelt, *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press, 1989.

[14] D. Montello, "The geometry of environmental knowledge," in *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*. Lecture Notes in Computer Science, vol. 639, A. Frank, I. Campari, and U. Formentini, Eds. Berlin, Germany: Springer-Verlag, 1992, pp. 136–152.

[15] M. Egenhofer and A. Shariff, "Metric details for natural-language spatial relations," *ACM Trans. Inf. Syst.*, vol. 16, no. 4, pp. 295–321, 1998.

[16] D. Hernández, *Qualitative Representation of Spatial Knowledge*. New York: Springer-Verlag, 1994.

[17] R. Grush, "Self, world and space: The meaning and mechanisms of ego- and allocentric spatial representation," *Brain Mind*, vol. 1, no. 1, pp. 59–92, Apr. 2000. (34).

[18] S. C. Levinson, "Frames of reference and Molyneux's question: Crosslinguistic evidence," in *Language and Space*, P. Bloom, M. Peterson, L. Nadel, and M. Garrett, Eds. Cambridge, MA: MIT Press, 1996, pp. 109–169.

[19] G. D. Logan and D. D. Sadler, "A computational analysis of the apprehension of spatial relations," in *Language and Space*, P. Bloom, M. A. Peterson, L. Nadel, and M. Garrett, Eds. Cambridge, MA: MIT Press, 1996, pp. 493–529.

[20] K. P. Gapp, "Angle, distance, shape, and their relationship to projective relations," in *Proc. 17th Conf. Cogn. Sci. Soc.*, 1995, pp. 112–117.

[21] T. Regier and L. Carlson, "Grounding spatial language in perception: An empirical and computational investigation," *J. Exp. Psychol.*, vol. 130, no. 2, pp. 273–298, 2001.

[22] H. Zimmer, H. Speiser, J. Baus, and A. Krüger, "Critical features for the selection of verbal descriptions for path relations," *Cognit. Process.*, 2001.

[23] P. E. Agre and D. Chapman, "What are plans for?" *Robot. Auton. Syst.*, vol. 6, no. 1/2, pp. 17–34, 1990.

[24] A. Aho, J. Hopcroft, and J. Ullman, *The Design and Analysis of Computer Algorithms*. Reading, MA: Addison-Wesley, 1974.

[25] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics: Principles and Practice in C*, 2nd ed. Reading, MA: Addison-Wesley, 1997.

[26] R. Power, "The organization of purposeful dialogues," *Linguistics*, vol. 17, pp. 107–152, 1979.

[27] D. Bikel, R. Schwartz, and R. Weischedel, "An algorithm that learns what's in a name," *Mach. Learn.*, vol. 34, no. 1–3, pp. 211–231, 1999.

[28] D. Roy, "Grounding words in perception and action: Computational insights," *Trends Cogn. Sci.*, vol. 9, no. 8, pp. 389–396, 2005.

[29] ——, "Semiotic schemas: A framework for grounding language in action and perception," *Artif. Intell.*, vol. 167, no. 1/2, pp. 170–205, 2005.

Michael Levit received the M.S. and Ph.D. degrees from the University of Erlangen, Erlangen, Germany, in 2000 and 2005, respectively.

He spent two years working with AT&T Laboratories, where he was actively engaged in the "How May I Help You?" project and has authored several scientific publications in the fields of spoken language recognition and understanding. In 2005, he was a Postdoctoral Associate with the MIT Media Laboratory. In 2006, he was with BBN Technologies, Cambridge, MA, and is currently with the International Computer Science Institute, Berkeley, CA, working on automatic question answering in multilingual environments.

Deb Roy received the B.A.Sc. degree in computer engineering from the University of Waterloo, Waterloo, ON, Canada, and the M.S. and Ph.D. degrees in media arts and sciences from the Massachusetts Institute of Technology (MIT), Cambridge.

He is an Associate Professor of media arts and sciences with MIT and Director of the Cognitive Machines Group at the MIT Media Laboratory. He has published papers in the areas of knowledge representation, speech and language processing, language acquisition, robotics, information retrieval, cognitive modeling, and human–machine interaction. He has served as Guest Editor for the journal *Artificial Intelligence* and is an Associate of the journal *Behavioral and Brain Sciences*.