Temporal Spatial Inverse Semantics for Robots Communicating with Humans

Ze Gong and Yu Zhang

***Abstract*— Effective communication between humans often embeds both temporal and spatial context. While spatial context captures the geographic settings of objects in the environment, temporal context describes their changes over time. In this paper, we propose *temporal spatial inverse semantics (TeSIS)* to extend the inverse semantics approach to also consider the temporal context for robots communicating with humans. Inverse semantics generates natural language requests while taking into account how well the human listeners would interpret those requests given the current spatial context. Compared to inverse semantics, our approach incorporates also temporal context by referring to spatial context information *in the past*.**

To achieve this, we extend the sentence structure in inverse semantics to generate sentences that can refer to not only the current but also previous states of the environment. A new metric based on the extended sentence structure is developed by breaking a single sentence into multiple independent sentences that refer to environment states at different times. Using this approach, we are able to generate sentences such as "*Please pick up the cup beside the oven that was on the dining table*". To evaluate our approach, we randomly generate scenarios in an experimental domain. Each scenario includes the description of the current and several immediate previous states. Natural language sentences are then generated for these scenarios using both inverse semantics that uses only the spatial context and our approach. Amazon MTurk is used to compare the sentences generated and results show that TeSIS achieves better accuracy, sometimes by a significant margin, than the baseline.

I. Introduction

Over the past few decades, we have seen robots gradually enter our lives, and they will undoubtedly continue to play a more and more important role in the future. Robots can provide assistance to humans in many areas, such as in household, industrial and medical applications. To improve efficiency and maintain safety during human-robot interaction, it is important for humans and robots to communicate with each other in an effective manner. While communication between humans are carried out naturally since we share a similar understanding of the world, the mismatch between humans and robots can be profound, and thus renders human-robot communication a significant challenge. For example, in natural language communication, the symbol grounding problem is related to the problem of how words get their meanings. An effective grounding process often embeds important context information to make communication more interpretable. However, differences in the understanding of the environment between humans and robots create a significant barrier for using such context information during communication. For example, a robot can easily refer to an object using its precise physical location, which however would be difficult for humans to understand.

In this paper, we focus on enabling robots to communicate information to humans using natural language sentences, or also known as the inverse symbol grounding problem. This is useful, for example, when robots need to request help from humans while performing a task, or convey specific information to them. In such situations, robots must generate natural language sentences that accurately communicate information to their human partners. Our research focus here is to enable robots to generate easily interpretable natural language sentences with context information.

The context information in natural language communication can be both spatial and temporal, which captures geographic settings of objects and their changes over time. For example, consider a warehouse like environment where a robot is helping a human with packing and shipping boxes. Everything is going on well until the robot hands over the human a box that has one item missing. The human is aware of the situation, adds the required item, and puts the box down on a table along with other boxes that are being packed. However, when the table is too full which makes it difficult for the robot to access the read-to-ship box, the robot must request help. Since there are other boxes on the table, the robot may use spatial context to specify which box it needs by asking "*Please hand me the red box that is on the right side of the table*". However, when the boxes are difficult to disambiguate using only the spatial context, it may be helpful to use the temporal context by asking "*Please hand me the box on the table that was just handed over to you*".

Tellex et al. [21] first introduced an approach called *inverse semantics* to address the inverse grounding problem for robots making requests to humans. Inverse grounding is considered within a grounding framework by simultaneously taking into account how well human listeners would interpret such requests given the current spatial context. First, to facilitate the understanding of natural language sentences for humans, the robot builds a mapping from words or phrases in a sentence to groundings in the environment. To enable robots to generate sentences, *inverse semantics* inverts this process by mapping from groundings in the environment to words or phrases that constitute sentences. One complexity in sentence generation is that instead of optimizing according to the human as a "speaker" (uttering commands to the robot), this inverse process must be optimized now for the human as a "listener" by reducing ambiguity in the communication. In simple scenarios, this approach can generate more accurate

Ze Gong and Yu Zhang are with the Computer Science and Engineering Department at Arizona State University, {zgong11, yzhan442}@asu.edu

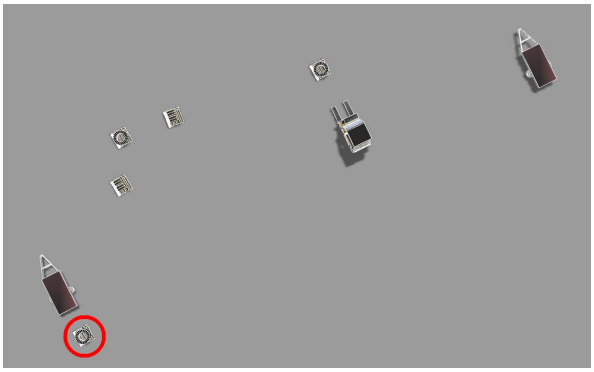
requests by extending sentences to include descriptive information about the spatial context. When the situation becomes more complicated, as we discussed, relying solely on spatial context becomes insufficient. We propose an approach called *temporal spatial inverse semantics (TeSIS)* that incorporates both spatial and temporal information that allows reference to spatial information in the past. Temporal context information is used whenever it helps resolve ambiguities, and thus TeSIS is expected to generate more interpretable sentences.

In this paper, we develop TeSIS to extend inverse semantics to incorporate temporal context for robots communicating with humans. First, we modify the context-free grammar (CFG) to generate sentences that can refer to both the current and previous environment states. A new metric is developed by breaking a single sentence into multiple independent sentences that map to environment states at different times. To evaluate our approach, we generate scenarios in a simulated experiment domain. Natural language requests are generated using *inverse semantics* and TeSIS. The generated sentences are evaluated on Amazon MTurk. Results show that TeSIS largely improves the accuracy of human performance for interpreting robot requests in a human-robot teaming setting.

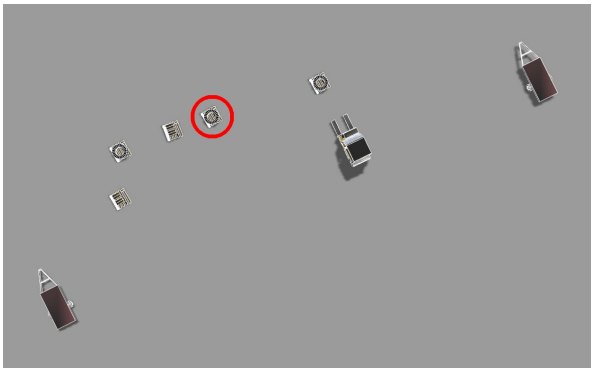
II. Related Work

Much research has been done for human-robot interaction using natural languages. For robots, this ability requires them to both understand natural language sentences [17], [18], [2], [22], [19] and be able to generate them [14], [20], [8], [16], [13]. Inverse semantics [21] used a probabilistic framework to model the process of natural language understanding [22] as an inference problem, and then invert this process for generating linguistic expressions. During the generation process, instead of choosing the sentence that best matches with the semantics, it selects the one that minimizes the ambiguity for humans by minimizing the uncertainty in the human understanding model. Our approach follows a similar process to facilitate human understanding of the sentence. The difference is that our approach generates sentences that also incorporate temporal context information from previous states, and hence can resolve ambiguities that cannot be easily addressed using only spatial context.

One of the important insights from [21] is that language understanding and generation are not two symmetrical processes [21]. In particular, in one direction, language understanding only needs to consider the most likely underlying semantics that a sentence is mapped to; in the other direction, however, language generation must not only consider how likely they are mapped, but also how likely they may be mis-mapped, in order to facilitate human understanding of the sentence. Such asymmetry not only exists in human-robot communication but also in many different research areas that involve human-robot interaction [11], such as robot motion planning [7], task planning [26], [6], learning and adaption [12], [3], and in general all aspects of robots decision making with humans in the loop [4]. To be seen as cooperative in such applications, robots are tasked with the additional burden to reduce human effort (usually at the cost of extra robot effort), whether physical or cognitive, via implicit [15] or explicit ways [10], [5].



(a) State at time $T-1$



(b) State at time T

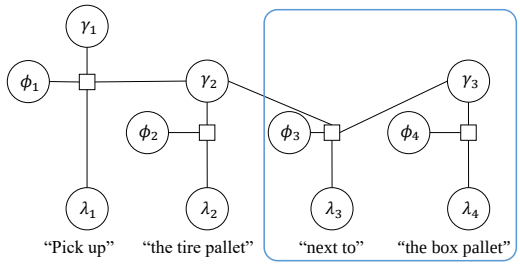
Fig. 1: A scenario in our experiment domain showing two contiguous time steps. This domain involves a robotic forklift and a human-operated forklift (not shown for clarity) to cooperatively load and unload pallets. The bottom part of this figure presents the four kinds of objects in this domain. The target object that the robot needs help with is marked in a red circle. (a) Environment state of the previous time step. (b) Environment state of the current time step.

III. Temporal Spatial Inverse Semantics

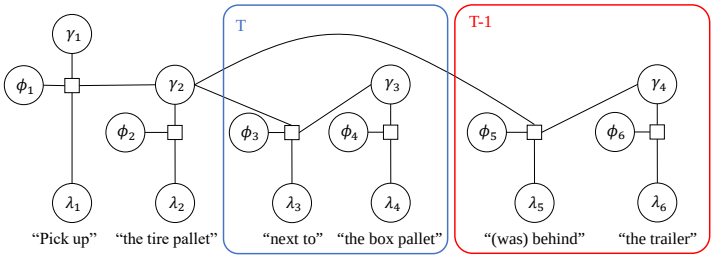
In this section, we describe our approach to temporal spatial inverse semantics, which extends *inverse semantics* [21] by incorporating temporal context information. First, we present our experiment domain, similar to that used in [21], which is also referred to throughout this section as a motivating scenario. Then, after a brief background discussion of inverse semantics, we introduce our extension.

A. Experiment Domain

Our experiment domain involves a robotic forklift and a human-operated forklift working together to load and unload pallets. Let us consider a scenario shown in Fig. 1, where there are two types of pallets, tire and box pallets. The tire



(a) G^3 generated by *inverse semantics* [21] for "*Pick up the tire pallet (that is) next to the box pallet*". The substructure in the blue bounding box is the extended part for spatial context.



(b) G^3 generated by *temporal spatial inverse semantics (TeSIS)* for "*Pick up the tire pallet (that is) next to the box pallet (which) was behind the trailer*". The first extended substructure in the blue bounding box maps phrases to groundings in the current time step T. The second extended substructure in the red bounding box maps phrases to groundings in the previous time step $T-1$.

Fig. 2: Examples of Generalized Grounding Graphs for *inverse semantics* [21] and our approach (*temporal spatial inverse semantics*). γ_i represents groundings in the environment. λ_i indicates constituent phrases from natural language commands. ϕ_i is a correspondence factor indicating the correctness of the mapping between groundings and constituent phrases.

pallet marked in a red circle is the target that the robot needs help with. Fig. 1a presents the state immediately before the current state; Fig. 1b presents the current state. For clarity, we are only showing the robotic forklift in the pictures. This scenario is difficult since the target tire pallet is located among other tire and box pallets. In this case, the sentence generated by *inverse semantics* that considers only spatial context may be "*(Please help me) Pick up the tire pallet that is next to the box pallets*". However, such an expression would be ambiguous for a human partner since there are two tire pallets next to the box pallets in the environment. While a more complex sentence may work, the sentence has to be meticulously structured in situations like this and thus becomes a challenge to understand. However, using temporal context can help immensely in this situation while keeping a simple sentence structure such as "*(Please help me) Pick up the tire pallet that was behind the trailer*". Thus, incorporating temporal context information from previous states can help reduce ambiguity and keep the sentence concise.

B. Generalized Grounding Graphs

Both *inverse semantics* and our *temporal spatial inverse semantics (TeSIS)* involve mapping between natural language phrases and groundings in the environment. We use Generalized Grounding Graphs (G^3) proposed by Tellex et al. [22], similar to that in inverse semantics. We will provide a short introduction about G^3 in this section.

G^3 is proposed for solving the problem of natural language understanding. Given a natural language command, such as "*Pick up the tire pallet*", the robot should be able to map the constituent phrases to groundings in the environment, shown in Fig. 2a. A grounding may be an object, place, or path in the environment, or an action that the robot can perform. Furthermore, a correspondence vector Φ is introduced for each phrase-grounding mapping. Each correspondence factor $\phi_i \in \Phi$ is used for indicating whether the mapping is right or wrong. To build a G^3, the natural language command is first decomposed into Spatial Description Clauses or SDCs introduced by Kollar et al. [16]. Then a G^3 is created according to the structural hierarchy of SDCs. The goal of natural language understanding for a sentence then becomes finding the groundings that maximize the following conditional probability:

$$\underset{\Gamma}{\operatorname{argmax}}\, p(\Phi = True|\text{command}, \Gamma) \tag{1}$$

where Γ represents all possible groundings in the environment. In other words, the system searches for a set of groundings that best match the command.

C. Inverse Semantics

Inverse semantics inverts the G^3 model to compute a mapping from groundings to a natural language sentence. For example, when the robot encounters a failure or an unexpected situation, it would generate a request to its human partner to help it restore to working condition and continue its execution. To achieve this, after associating the groundings of a desired action to a G^3 model, the system searches for a sentence that would best convey the request to the human based on the language understanding model. A metric h is used to estimate the quality of a generated natural language sentence, which must be maximized:

$$\underset{\Lambda}{\operatorname{argmax}}\, h(\Lambda, a, M) \tag{2}$$

The environmental context M contains location, orientation and path information of each object. a is the desired action that the robot expects the human to help with. The aim is to find an utterance Λ that maximizes h above. An intuitive method to specify h is to consider a sentence that corresponds to the groundings of the desired action without considering other possible groundings of the sentence in the environment. This however would work only in simple environments. In more complex environments, it is bound to lead to ambiguities. For example, when there are two tire pallet in the environment, a request of "*Pick up the tire pallet*" would fail to convey the intended meaning and lead to confusion.

$S \rightarrow VB\ NP$
$S \rightarrow VB\ NP\ PP$
$PP \rightarrow IN\ NP$
$NP \rightarrow NP\ VBD\ PP$
$NP \rightarrow NP\ VBD\ VB\ PP$
$VB \rightarrow$ pick up | load | unload | drive forward
$NP \rightarrow$ the trailer | the forklift | the tire pallet
the box pallet | the trailer | you
$IN \rightarrow$ on | to the left of | to the right of
behind | in front of | from | to
$VBD \rightarrow$ was | were

Fig. 3: Part of the context-free grammar that generates the search space for natural language sentences (commands) for our experiment domain.

To address this problem, the key is to take the model of the listener into consideration in the metric h. First, the candidate sentence structure (based on the CFG for generating natural language requests) is extended to allow more context information to be incorporated into the request. After generating the candidate sentences, all the possible groundings are considered in the environment for each candidate. The probability that the human will correctly interpret the requested action given a candidate sentence is measured as follows:

$$h(\Lambda, \gamma_a^*, M) = \frac{\sum_{\Gamma|\gamma_a=\gamma_a^*} p(\Phi|\Gamma, \Lambda, M)}{\sum_{\Gamma'} p(\Phi|\Gamma', \Lambda, M)} \quad (3)$$

where γ_a denotes all grounding variables that are associated with the desired action and γ_a^* the desired action groundings. This h metric takes into account human interpretation of a given natural language request by minimizing the uncertainty in the distribution of groundings that are consistent with γ_a^*, thus minimizing potential confusion due to other possible groundings for the sentence.

D. Temporal Spatial Inverse Semantics

By modeling both the listener and environment, *inverse semantics* performs better than a baseline that just models the environment (through the phrase-grounding mapping). The approach is able to generate sentences like "*Pick up the tire pallet in front of the forklift*" instead of simply "*Pick up the tire pallet*". However, as we mentioned, while this approach works fine when the spatial context is sufficient to remove ambiguity, it may fail to generate effective communication in a complex environment that is rich of similar spatial features, e.g., environments that contain multiple objects of the same type such as a warehouse. In such cases, we need a more powerful approach that can incorporate more context information to help resolve ambiguity.

To achieve this, our system uses also the temporal context by considering previous environment states when generating natural language requests. Hence, we call our approach *temporal spatial inverse semantics (TeSIS)*. First, we extend the CFG sentence generator for generating candidate sentences, shown in Fig. 3. The extended structure allows references to spatial context from both the current and previous states. The generated candidates are then fit into a G^3 model. Fig. 2a shows an example of a G^3 graph for inverse semantics. Fig. 2b shows an example for *TeSIS* for a model that considers the current state and a previous state. There are two parts in this graph. One part is for mappings between phrases and groundings in the current time step (similar to inverse semantics); the other part is the extended part for mappings between phrases and groundings in the previous state.

A new metric h' that incorporates the consideration of context information from previous states (up to n) can be specified as follows:

$$h'(\Lambda_{0,k}, \gamma_a^*, M_{0,k}) = p\left(\gamma_{0,k}^* | \Phi_{0,k}, \Lambda_{0,k}, M_{0,n}\right) \quad (4)$$

where $\gamma_{0,k}^*$ represents $\gamma_{a_0} = \gamma_a^*, \gamma_{o_1} = \gamma_o^*, \cdots, \gamma_{o_k} = \gamma_o^*$, which correspond to references into the current and previous states (indexed from $0-k$ with 0 being the current state), respectively. The robot will maintain the last n environment states from the previous time steps. The extended sentence structure (with k expanded substructures referring to the previous states) of a new sentence maps constituent phrases to groundings in k out of the n previous states. The action grounding variables γ_{a_0} in the current state correspond to the desired action groundings γ_a^*. For previous time steps, we track the target object that the robot expects the human to perform the action on. We denote the grounding of this object as γ_{o_i} in the ith substructure and the desired grounding value as γ_o^* (i.e., the target). Therefore, given the constituent phrases and correspondence vectors of a sentence's G^3 model and all the context information required (i.e., about the previous states), we can now compute the quality metric h' that captures how a human would correctly interpret the robot's request as follows.

Out of the n previous environment states, there are C_n^k possible combinations using k previous states. We compute a quality measure (the equation inside the summation in Eq. (5) that captures the uncertainty in the correct mapping) for each possible combination. Assuming that all combination are equally likely, we then take their average as h'. In the following, we refer to the set of all C_n^k combinations as S_k:

$$h'(\Lambda_{0,k}, \gamma_a^*, M_{0,k}) = \frac{1}{|S_k|} \sum_{MC_k \in S_k} p\left(\gamma_{0,k}^* | \Phi_{0,k}, \Lambda_{0,k}, MC_k\right) \quad (5)$$

where MC_k represents the environment context that corresponds to a combination of the current and previous k states. For computing the quality measure for each combination, we marginalize over values of all groundings from Γ_0 to Γ_k (denoted as $\Gamma_{0,k}$):

$$h'(\Lambda_{0,k}, \gamma_a^*, M_{0,k}) = \frac{1}{|S_k|} \sum_{MC_k \in S_k} \sum_{\Gamma_{0,k}^*} p\left(\Gamma_{0,k} | \Phi_{0,k}, \Lambda_{0,k}, MC_k\right) \quad (6)$$

where $\Gamma_{0,k}^*$ is used for representing $\Gamma_0|\gamma_{a_0} = \gamma_a^*, \Gamma_1|\gamma_{o_1} =$

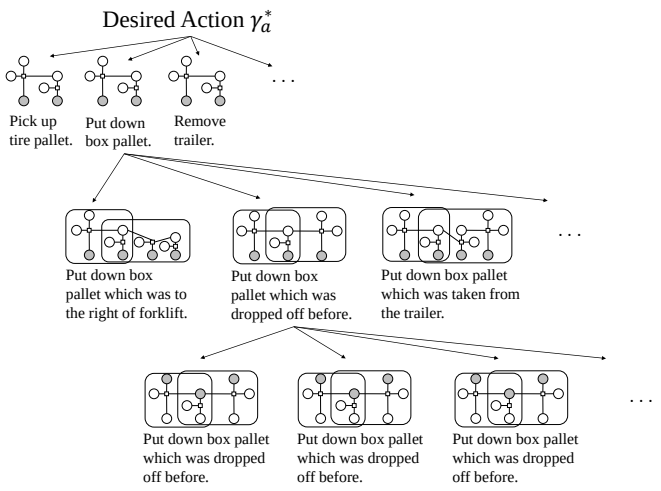


Fig. 4: Search schema for *temporal spatial inverse semantics (TeSIS)*. At level 1, given the desired action groundings γ_a^*, we search over the predefined natural language corpus to find the top m sentence candidates. At the second level, sentences are extended by using the CFG sentence generator that allows references to k selected previous environment states. At the third level, we search over all the groundings from both the previous and current environment states to find the sentence with the highest value for h'.

$\gamma_o^*, \cdots, \Gamma_k | \gamma_{o_k} = \gamma_o^*$. Apply Bayes' rule to factor the model:

$$h'(\Lambda_{0,k}, \gamma_a^*, M_{0,k}) = \frac{1}{|S_k|} \cdot \sum_{MC_k \in S_k} \sum_{\Gamma_{0,k}^*} \frac{p\left(\Phi_{0,k} | \Gamma_{0,k}, \Lambda_{0,k}, MC_k\right) p\left(\Gamma_{0,k} | \Lambda_{0,k}, MC_k\right)}{p\left(\Phi_{0,k} | \Lambda_{0,k}, MC_k\right)} \quad (7)$$

Then we rewrite the denominator as the summation over all the groundings from Γ_0' to Γ_k':

$$h'(\Lambda_{0,k}, \gamma_a^*, M_{0,k}) = \frac{1}{|S_k|} \cdot \sum_{MC_k \in S_k} \sum_{\Gamma_{0,k}^*} \frac{p\left(\Phi_{0,k} | \Gamma_{0,k}, \Lambda_{0,k}, MC_k\right) p\left(\Gamma_{0,k} | \Lambda_{0,k}, MC_k\right)}{\sum_{\Gamma_{0,k}'} p\left(\Phi_{0,k} | \Gamma_{0,k}', \Lambda_{0,k}, MC_k\right) p\left(\Gamma_{0,k}' | \Lambda_{0,k}, MC_k\right)} \quad (8)$$

The summation in the denominator is for all the groundings in the k selected environment states while the summation outside is for the action or object groundings in each selected state. We therefore can move the summation to the nominator:

$$h'(\Lambda_{0,k}, \gamma_a^*, M_{0,k}) = \frac{1}{|S_k|} \cdot \sum_{MC_k \in S_k} \frac{\sum_{\Gamma_{0,k}^*} p\left(\Phi_{0,k} | \Gamma_{0,k}, \Lambda_{0,k}, MC_k\right) p\left(\Gamma_{0,k} | \Lambda_{0,k}, MC_k\right)}{\sum_{\Gamma_{0,k}'} p\left(\Phi_{0,k} | \Gamma_{0,k}', \Lambda_{0,k}, MC_k\right) p\left(\Gamma_{0,k}' | \Lambda_{0,k}, MC_k\right)} \quad (9)$$

Similar to [21], we assume that Γ and Λ are independent when we do not know Φ. $p\left(\Gamma_{0,k} | \Lambda_{0,k}, MC_k\right)$ then becomes a constant and we can remove it from both the nominator and denominator, which gives:

$$h'(\Lambda_{0,k}, \gamma_a^*, M_{0,k}) = \frac{1}{|S_k|} \sum_{MC_k \in S_k} \frac{\sum_{\Gamma_{0,k}^*} p\left(\Phi_{0,k} | \Gamma_{0,k}, \Lambda_{0,k}, MC_k\right)}{\sum_{\Gamma_{0,k}'} p\left(\Phi_{0,k} | \Gamma_{0,k}', \Lambda_{0,k}, MC_k\right)} \quad (10)$$

Here $p\left(\Phi_{0,k} | \Gamma_{0,k}, \Lambda_{0,k}, MC_k\right)$ is factorized as follows as the multiplication of factor values in the entire G^3 structure from all the time steps we refer to:

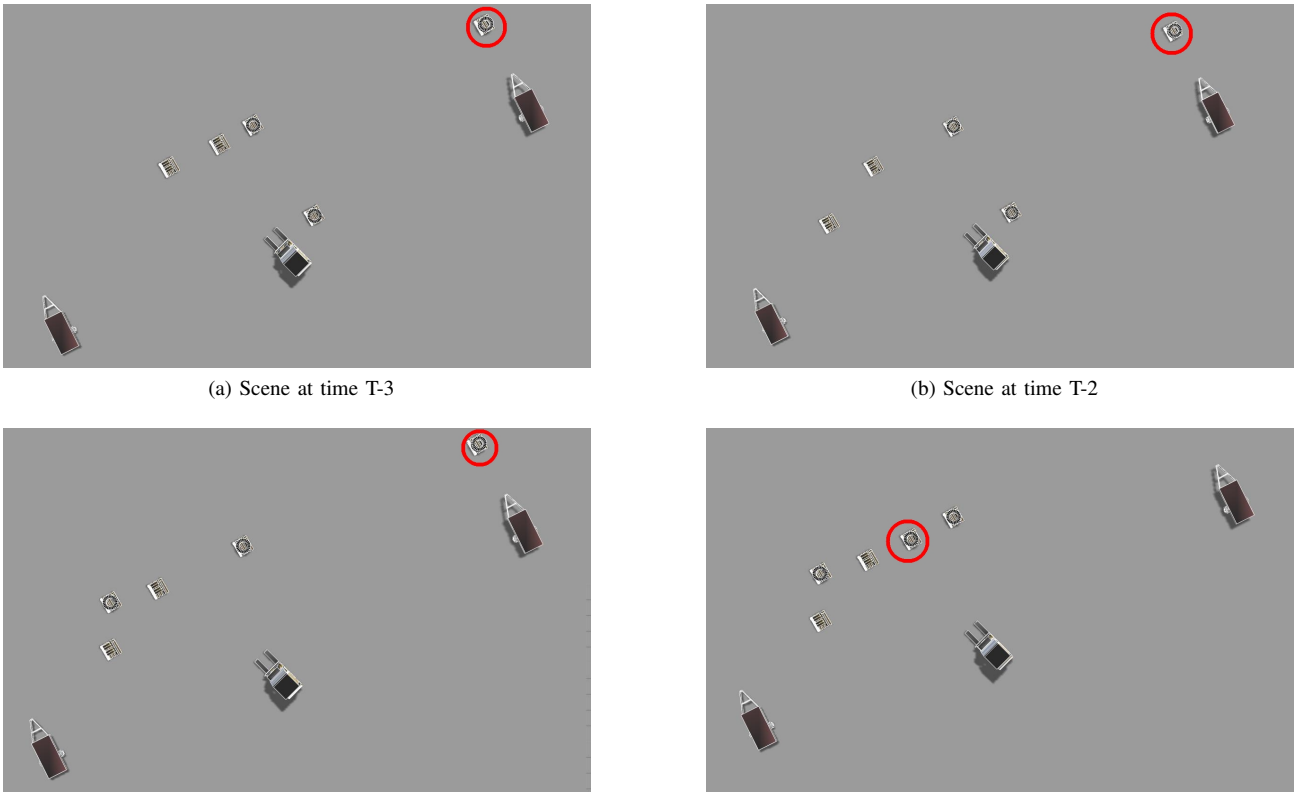
$$p\left(\Phi_{0,k} | \Gamma_{0,k}, \Lambda_{0,k}, MC_k\right) = \frac{1}{Z} \prod_i \Psi_i\left(\phi_i, \lambda_i, \gamma_i, M^{(i)}\right) \quad (11)$$

where Z is the normalization constant and $M^{(i)}$ is the environmental state that the ith factor lies in. Equation (10) shows the new evaluation metric h' for *temporal spatial inverse semantics*. We compute the quality measure for each possible combination of k previous environmental states and then take the average of them to obtain the quality of the candidate sentence. By generating different sentence structures that may or may not refer to the previous states, *temporal spatial inverse semantics* extends sentence structures to use temporal context only when it helps with conveying the information.

The search process of generating natural language requests for the robot in TeSIS, however, is computationally expensive. Hence, heuristics are adopted to make it practical. Greedy decisions are made at the three different levels, respectively, for word & phrase selection, sentence structure extension, and groundings exploration. The work flow of TeSIS is shown in Fig. 4, where levels from top down, respectively, correspond to word & phrase selection, structure extension, and grounding exploration. At the first (top) level, we first convert the desired action a to the corresponding groundings γ_a^*, which allows us to create a G^3 model for γ_a^*. Given the desired groundings and the G^3 structure, we search over the predefined corpus of natural language words and phrases to find the top m candidate sentences. Next we extend the structure of candidate sentences from level 1 (top level in Fig. 4) by the CFG shown in Fig. 3. This extension includes extending substructures for all k previous environment states. At the last level, all the groundings from the k previous environment states and current state are explored for computing the value of h'. At each level, only selected candidates are passed to the next level. This greedy search process significantly reduces the number of sentences and groundings to be considered.

IV. Evaluation

We evaluate our approach in the forklift domain that was adapted from the one used in [22]. For a given scenario, the robotic forklift would be executing actions following a given plan. During its plan execution, it may encounter unexpected situations such that the execution has to stop. We consider unexpected situations as prerequisites that must be met and however cannot be performed by the robot. In such cases, the robot must generate requests to solicit help from the human who is also operating a forklift nearby.



(a) Scene at time T-3

(b) Scene at time T-2

(c) Scene at time T-1

(d) Scene at time T

Fig. 5: Test scenario with the current time step and three previous time steps. The target object is in the red circle. Generated requests: (1) G^3 inverse semantics: "*Pick up (the) tire pallet to the right of (the) forklift*"; (2) G^3 temporal spatial inverse semantics with $k = 2$ and $n = 3$: "*Pick up (the) tire pallet which was in front of (a) trailer*"; (3) G^3 temporal spatial inverse semantics with $k = 1$ and $n = 3$: "*Pick up (the) tire pallet which was in front of (a) trailer*."

A. Training and Testing

We adapted the dataset from [1] for training and testing in a forklift domain for manipulation and navigation. There are 22 different scenarios that include loading, unloading and moving from one location to another location. Each scenario contains 13 human written commands on average describing actions the forklift performs. We trained the model of natural language understanding following the same procedure as in Tellex et al. [22].

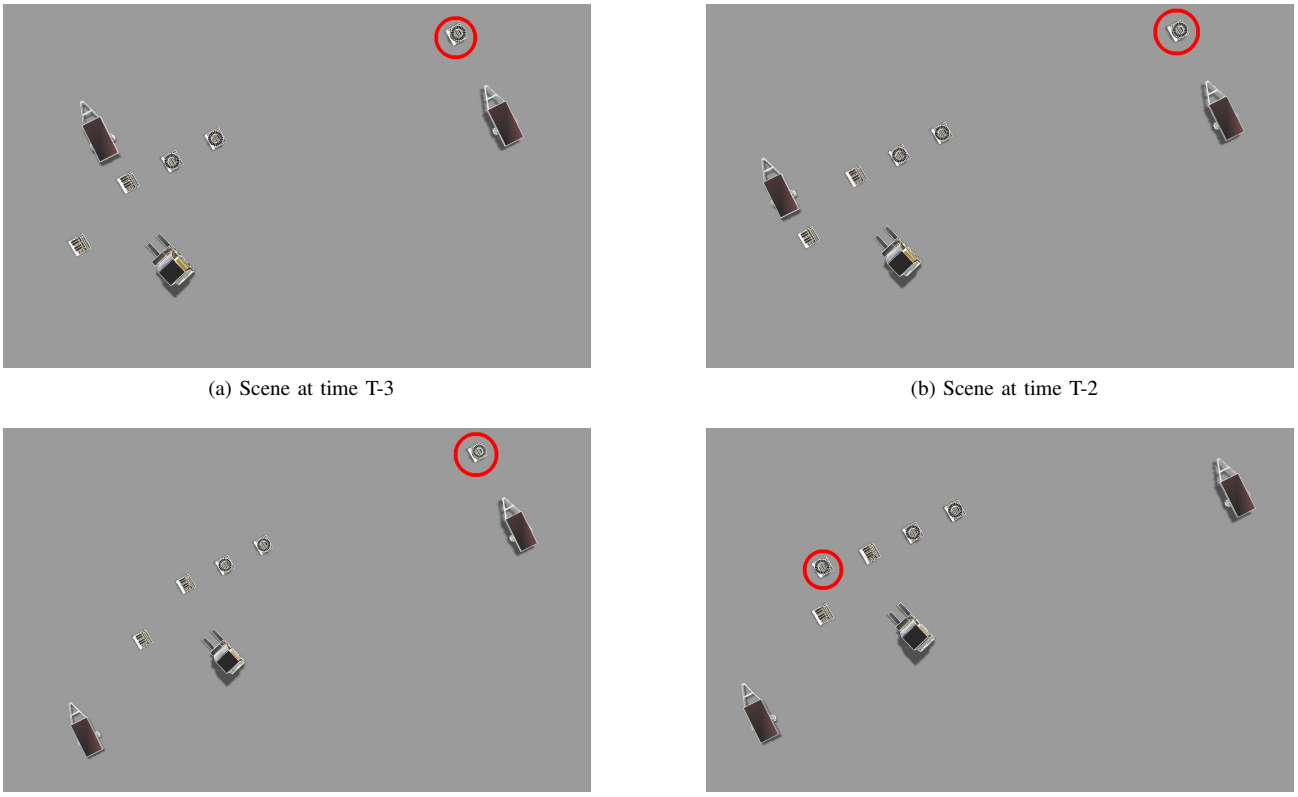
For testing, we must first create scenarios that span across multiple time steps. Given each scenario in the training dataset, we first randomly select an object in the environment, and modify its grounding information, such as its location and orientation, to create an environment state for the previous time step. To obtain environment states for multiple previous time steps, we follow the same procedure iteratively. Meanwhile we check the generation process to avoid collision among objects. We created 83 test scenarios. Each one contains 4 environment states in total, including the current state (i.e., 3 previous states).

B. Evaluation and Discussion

To evaluate our system, we use Amazon MTurk (AMT). First, for each test scenario, we create a scene in the simulator as in Fig. 1 for each environment state. Each scenario involves two trailers, five pallets with two different types (see 1). We present the workers on AMT all the 4 scenes in order with a generated natural language request together. The turkers are asked to choose from multiple choices the object that the request refers to. This provides a direct measurement for how accurate the generated requests are interpreted.

We experiment with three experiment settings. One of the settings uses *inverse semantics* that considers only spatial context, the other two settings both use *temporal spatial inverse semantics*, one with $k = 2$ and $n = 3$, and the other with $k = 1$ and $n = 3$. The aim here is to also shed light on how much temporal context humans are accustomed to. Table I presents the results for all the three settings; each setting used 100 turkers; each turker responded to about 30 scenarios. Both settings that use *temporal spatial inverse semantics* perform better than *inverse semantics*. In some scenarios where spatial context can be used to disambiguate well, *inverse semantics* with S_2 does a pretty good job. For example, when the box pallet that is referred to is exactly located to the left of the trailer and all the other pallets are on the other side of the trailer. Nevertheless, generally, we can see that these scenarios are already getting difficult for inverse semantics that uses only spatial context.

Sometimes, obviously ambiguous spatial description may be used by inverse semantics. One example that we observed



(a) Scene at time T-3

(b) Scene at time T-2

(c) Scene at time T-1

(d) Scene at time T

Fig. 6: Test scenario with the current time step and three previous time steps. The target object is marked in the red circle. Generated requests: (1) G^3 inverse semantics: "*Pick up (the) tire pallet in front of (the) box pallet*"; (2) G^3 temporal spatial inverse semantics with $k = 2$ and $n = 3$: "*Pick up (the) tire pallet which was to the left of (a) trailer (and had been) in front of (the) trailer*"; (3) G^3 temporal spatial inverse semantics with $k = 1$ and $n = 3$: "*Pick up (the) tire pallet which was in front of (a) trailer*".

from the test scenarios is shown in Fig. 5. In the current state, the target tire pallet is located among other pallets. The top 1 sentence generated by *inverse semantics* is "*Pick up (the) tire pallet to the right of forklift*". This however is likely to cause much confusion. In contrast, observing the three previous states, we can see that the target tire pallet was in front of the trailer on the right. In such cases, incorporating temporal context information from previous time steps would helps immensely to avoid such ambiguity. Indeed, the top sentence generated by both *temporal spatial inverse semantics* settings are "*Pick up (the) tire pallet which was in front of (a) trailer*" For *temporal spatial inverse semantics* with $k = 2$, the two extended substructures that correspond to the two previous state may be mapped to the exact same constituent phrases. In such case, we trim the sentences to include only one extended substructure to avoid the repetition.

Comparison between the two settings using *temporal spatial inverse semantics* shows that generating sentences by referring to only one previous state out of multiple available states seems to perform better than referring to multiple previous states. This result may be explained in multiple different ways. First, consider the implications of adding temporal context. To take advantage of the previous state information, we compute the quality measures for each k combination given the extended sentence and its corresponding G^3 structure. Then, the top sentence is selected by picking the highest value of h' which is computed as the average of these quality measures. When $k = 1$, we will compute the quality measure given each previous state that may be selected one by one, and then return the value of h' and select the top sentence. For $k = 2$, we must consider all the combinations of the different previous two time steps that may be selected. Given that heuristics are used, the more complex the inference problem is, the most likely that inaccuracy may be introduced.

One example for illustrating this effect is presented in Fig. 6. Similar to the test scenario shown in Fig. 5, the target tire pallet is located among other pallets in the current state and it was in front of the trailer on the right in the previous states. The top sentence generated by *temporal spatial inverse semantics* with $k = 2$ is "*Pick up (the) tire pallet which was to the left of (a) trailer (and had been) in front of (the) trailer*", and with $k = 1$ "*Pick up the tire pallet which was in front of trailer*". Both settings successfully capture the temporal context in the previous time steps which specify that the tire pallet was in front of the right trailer. The sentence generated when $k = 2$, however, includes a somewhat confusing description in the extended substructure

TABLE I: Comparison of the Successful Interpretation Rates of the Generated Requests

Metric	% Success
G^3 inverse semantics	56.25
G^3 temporal spatial inverse semantics with $k = 2$ and $n = 3$	58.33
G^3 temporal spatial inverse semantics with $k = 1$ and $n = 3$	66.67

that refers to a second previous state. Of course, another simpler explanation is that humans prefer shorter sentences than longer sentences.

V. Conclusion

In this paper, we introduce an approach called *temporal spatial inverse semantics (TeSIS)* for generating unambiguous natural language sentences for robots communicating with humans. Rather than only using spatial context, we also take into account temporal context by considering environment states from previous time steps to further help resolve ambiguities. First, we extend the sentence structure specified as a CFG to allow references to the previous states. A new metric based on the extended sentence structures is developed that breaks a single sentence to multiple independent sentences that can be inferred separately. In such a way, our approach generates sentences that can capture context information in the past whenever it helps convey the information. Our evaluation demonstrates promising performance of TeSIS for generating unambiguous requests for human partners to understand.

Although TeSIS presents promising results, occasionally, when even the combination of spatial and temporal context is insufficient, generating requests would become very challenging. In such cases we can use an interactive approach with humans in the loop as in [24], by engaging in a dialog based conversation to resolve ambiguities as in [23]. However, such communication has a high requirement of human cognition. Alternatively, it is also possible to study multi-modal communication to generate natural language, gesture, visual, and other communication cues [9] at the same time. We also plan to use TeSIS for generating signaling [10] in human-robot interaction. This will be combined with our approaches to generating explicable plans [26], [25] and explanations [5], [4] to faciliate fluent human-robot teaming.

Acknowledgment

We would like to thank Dr. Stefanie Tellex for providing the code for modeling natural language command understanding [22] and the dataset [1] that was adapted in this work. This research is supported in part by the NASA grant NNX17AD06G.

References

[1] [accessed september 15, 2017] spatial language understanding framework. [Online]. Available: http://people.csail.mit.edu/stefie10/slu/

[2] D. Arumugam, S. Karamcheti, N. Gopalan, L. L. Wong, and S. Tellex, "Accurately and efficiently interpreting human-robot instructions of varying granularities," *arXiv preprint arXiv:1704.06616*, 2017.

[3] H. Ben Amor, G. Neumann, S. Kamthe, O. Kroemer, and J. Peters, "Interaction primitives for human-robot cooperation tasks," in *Proceedings of 2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[4] T. Chakraborti, S. Kambhampati, M. Scheutz, and Y. Zhang, "Ai challenges in human-robot cognitive teaming," *arXiv preprint arXiv:1707.04775*, 2017.

[5] T. Chakraborti, S. Sreedharan, Y. Zhang, and S. Kambhampati, "Explanation generation as model reconciliation in multi-model planning," *IJCAI*, 2017.

[6] T. Chakraborti, Y. Zhang, D. Smith, and S. Kambhampati., "Planning with resource conflicts in human-robot cohabitation," in *AAMAS*, 2016.

[7] A. Dragan and S. Srinivasa, "Generating legible motion," in *Proceedings of Robotics: Science and Systems*, Berlin, Germany, June 2013.

[8] R. Fang, M. Doering, and J. Y. Chai, "Collaborative models for referring expression generation in situated dialogue." in *The Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI, 2014, pp. 1544–1550.

[9] ——, "Embodied collaborative referring expression generation in situated human-robot interaction," in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2015, pp. 271–278.

[10] Z. Gong and Y. Zhang, "Robot signaling its intentions in human-robot teaming," in *HRI Workshop on Explainable Robotic Systems*, 2018.

[11] S. Guadarrama, L. Riano, D. Golland, D. Go, Y. Jia, D. Klein, P. Abbeel, T. Darrell, *et al.*, "Grounding spatial relations for human-robot interaction," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2013*. IEEE, 2013, pp. 1640–1647.

[12] D. Hadfield-Menell, S. J. Russell, P. Abbeel, and A. Dragan, "Cooperative inverse reinforcement learning," in *Advances in neural information processing systems*, 2016, pp. 3909–3917.

[13] T. M. Howard, I. Chung, O. Propp, M. R. Walter, and N. Roy, "Efficient natural language interfaces for assistive robots," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2014*. IEEE, 2014.

[14] D. Jurafsky and J. H. Martin, *Speech and language processing*. Pearson London:, 2014, vol. 3.

[15] R. A. Knepper, C. I. Mavrogiannis, J. Proft, and C. Liang, "Implicit communication in a joint action," in *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 2017, pp. 283–292.

[16] T. Kollar, S. Tellex, D. Roy, and N. Roy, "Toward understanding natural language directions," in *Proceedings of the 5th ACM/IEEE international conference on Human-robot interaction*. IEEE Press, 2010, pp. 259–266.

[17] C. Matuszek, E. Herbst, L. Zettlemoyer, and D. Fox, "Learning to parse natural language commands to a robot control system," in *Experimental Robotics*. Springer, 2013, pp. 403–415.

[18] D. K. Misra, J. Sung, K. Lee, and A. Saxena, "Tell me dave: Context-sensitive grounding of natural language to manipulation instructions," *The International Journal of Robotics Research*, vol. 35, no. 1-3, pp. 281–300, 2016.

[19] R. Paul, J. Arkin, N. Roy, and T. M. Howard, "Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators." in *Robotics: Science and Systems*, 2016.

[20] E. Reiter and R. Dale, *Building natural language generation systems*. Cambridge university press, 2000.

[21] S. Tellex, R. A. Knepper, A. Li, D. Rus, and N. Roy, "Asking for help using inverse semantics." in *Robotics: Science and Systems*, vol. 2, no. 3, 2014.

[22] S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. J. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation." in *The Twenty-Fifth AAAI Conference on Artificial Intelligence*, vol. 1. AAAI, 2011, p. 2.

[23] S. Tellex, P. Thaker, R. Deits, D. Simeonov, T. Kollar, and N. Roy, "Toward information theoretic human-robot dialog," *Robotics*, p. 409, 2013.

[24] J. Thomason, J. Sinapov, M. Svetlik, P. Stone, and R. J. Mooney, "Learning multi-modal grounded linguistic semantics by playing "i spy"." in *IJCAI*. IJCAI, 2016, pp. 3477–3483.

[25] Y. Zhang, S. Sreedharan, A. Kulkarni, T. Chakranaroti, H. H. Zhuo, and S. Kambhampati, "Plan explicability for robot task planning," in *RSS Workshop on Planning for Human-Robot Interaction*, 2016.

[26] ——, "Plan Explainability and Predictability for Robot Task Planning," in *ICRA*, 2017.