# Everything Robots Always Wanted to Know about Housework (But were afraid to ask)

Daniel Nyga[1] and Michael Beetz[1,2]
[1]Intelligent Autonomous Systems Group, Technische Universität München, Germany
[2]Research Group Artificial Intelligence, University of Bremen, Germany
{nyga, beetz}@cs.tum.edu

***Abstract*— In this paper we discuss the problem of action-specific knowledge processing, representation and acquisition by autonomous robots performing everyday activities. We report on a thorough analysis of the household domain, which has been performed on a large corpus of natural-language instructions from the Web and underlines the supreme need of action-specific knowledge for robots acting in those environments. We introduce the concept of *Probabilistic Robot Action Cores (PRAC)* that are well-suited for encoding such knowledge in a probabilistic first-order knowledge base. We additionally show how such a knowledge base can be acquired by natural language and we address the problems of incompleteness, underspecification and ambiguity of naturalistic action specifications and point out how PRAC models can tackle those.**

## I. Introduction

Roadmaps for robot research and technology identify robotic (co-)workers, assistants, and companions as promising targets for our technologies. These applications have in common that autonomous robots must perform complete jobs including a variety of human-scale activities, can be tasked in naturalistic ways, and operate over extended periods of time. They have to perform what is commonly referred to as *everyday activity*. In a recent experiment, Beetz *et al.* [1] have shown the feasibility of such robotic assistants by executing a natural-language recipe for making pancakes on one of their robotic platforms, which is illustrated in Figure 1.

Anderson [2] defines an everyday activity as a complex task that is common and mundane to the human – in our case to the robot – and that it in turn has a *great deal of knowledge* about. People performing everyday activity typically aim for adequate or satisficing performance rather than optimality and expert performance. Following this definition, everyday activities are routine in a sense that they occur frequently and, as a consequence, a robot must be well-experienced in performing them. The familiarity with these activities hence results in a large amount of knowledge about *how* a particular activity is to be performed in a specific context, going along with the awareness about objects involved in these actions as well as their inter- and intrarelations. This makes executing everyday activities a very knowledge-intensive task rather than intensive to planning.

As a consequence, knowledge about actions and objects serves as a source for constraints, i.e. a robot is to perform a particular action subject to the constraints given by its knowledge about this action. Constraining the space of possible action configurations by knowledge does not mean to restrict the physical or mental capabilities of a robot, but to guide the process of decision making. Considering only alternatives that have worked before rather rules out alternatives that are inappropriate or irrational in a particular context and hence makes repeated reasoning about choices obsolete. Consider, for example, a naturalistic instruction such as "Push the spatula under the pancake," which might be taken from a recipe for making pancakes. Successfully executing this instruction requires the robot to *hold* the spatula at its handle, to push only the blade of the spatula under the pancake, to push the blade at a position where the pancake can be lifted safely, to push the blade between the pancake and the pan, and to hold the spatula in an appropriate angle to the pan, to name only a few. As these additional constraints are not stated explicitly in the original instruction, the robot has to infer them by itself. This requires us to equip robots with a substantial body of knowledge that enables it to resolve ambiguities and to infer information that is missing in naturalistic task specifications.


Fig. 1. The robot "TUM-Rosie" while making pancakes. The pancake recipe has been downloaded from the web page wikihow.com.

Knowledge allows a robot to adjust action parameters such as the object acted on, the utensil or tool used to manipulate the object, the direction and destination of an action etc. or, in other words, to perform the *appropriate* action on the *appropriate* objects in an *appropriate* way. Appropriateness here does *not* mean optimality. Simon [3] refers to this

"satisficing" performance in decision making as *rational boundedness*, which results from cognitive limitations of an agent that has to balance utility and costs of deliberation effort. In these terms, an optimal solution often cannot be found in real-world scenarios and a robot "accepts 'good enough' alternatives, not because he prefers less to more but because he has no choice." [3]

These considerations set the stage for this paper. We intend to investigate *how much action-specific knowledge* robots should be equipped with in order to be capable of performing everyday activities successfully, effectively and competently. We will further investigate how this knowledge can be *acquired*, *represented* and *used*. For this research we restrict ourselves to abstract knowledge, the knowledge about actions that is typically communicated and written down in instructions. Other kinds of essential knowledge that the abstract knowledge needs to be combined with is being addressed in other works and includes naive physics knowledge [4], common-sense knowledge [5] and knowledge gathered from experience [6].

The contributions of this paper are the following:

- We will hypothesize and give estimations for the amount of knowledge that robots might need for the competent performance of envisioned robot jobs such as preparing meals. These estimations will be obtained by the application of data mining techniques to websites that provide written instructions for performing everyday activities that are intended for human use.
  Based on these results we will discuss (1) how this knowledge is organized, (2) how it is to be represented and (3) how it is to be acquired.
- We will introduce the concept of *Probabilistic Robot Action Cores* (*PRAC*) for representing action-specific knowledge for everyday manipulation, which can be thought of as abstract event types encoded in a probabilistic first-order knowledge base.
- We will finally demonstrate the strength of PRAC models by applying a proof-of-concept implementation to the problems of disambiguation and completion of naturalistic action specifications.

## II. How much Action-specific Knowledge Does a Robot Need?

When building a domain-specific knowledge base (KB), it is crucial to have a thorough understanding of the requirements the KB has to meet. In order to get an idea about the knowledge required for executing everyday activities, we did a study on a large set of natural-language instructions that we mined from the `wikihow.com` website, which contains thousands of recipes, plans and other step-by-step directives for a vast number of everyday household activities, which are written by humans and intended for human use. Our investigations aim at answering the following questions:

- *How many different actions does a robot need to know?*
- *How much variation do these actions exhibit?*
- *How do humans explain complex tasks to other humans?*

| ***Action Verb*** | ***# Occurrences*** |
|---|---|
| Adding sth. to sth./Combining | > 7,900 |
| Picking/Placing sth. swh. | > 4,900 |
| Filling/Pouring sth. into/onto sth. | > 3,100 |
| Removing sth. | > 1,700 |
| Stirring/Beating sth. | > 1,900 |
| Serving sth. | > 1,400 |
| Mixing/Blending sth. | > 1,200 |

Fig. 2. Most frequent action verbs in the wikiHow.com dataset and their number of occurrences. Further frequent action verbs comprise Baking, Cooking/Simmering/Boiling, Cutting/Chopping/Slicing, Sprinkling, Flipping/Turning over, Refrigerating/Cooling/Freezing, Shaking, Waiting.

Especially the third question is of particular importance in order to be able to endow robots with means for understanding naturalistic task specifications.

At the time this study has been conducted, the "Food & Entertaining" category on `wikiHow.com` comprises 273 subcategories consisting of 8786 articles in natural language, covering basic cooking skills (e.g. cutting techniques), recipes for cooking a wide range of dishes (e.g. making pancakes) or plans for organizing a whole dinner party.

The plans have been automatically extracted by a website parser. For analyzing them, we applied the system described in [7] for extracting single instructions from natural-language text and for transforming them into a formal, logic-based representation. The system first parses a natural-language text using a statistical parser and afterwards exploits its syntactic structure in order to determine action verbs as well as semantic relations such as the object acted on and prepositional relations. The system also implements the mapping from word meanings to concepts in a logical knowledge base. For a detailed description we refer to [7].

### A. How many actions are there?

We analyzed more than 130,000 sentences out of which we extracted about 53,000 relevant instructions. Here, instructions are regarded as relevant if they are goal-directed in a sense that a respective instruction effectively contributes to the overall outcome of a plan. This differentiation is important since in recipes, beside regular instructions, also a vast amount of additional explanations, comments and non-goal-directed instructions such as "Enjoy your meal," "Admire your work," or "Be an artist," for instance, can be found. Such instructions do not represent action verbs for object manipulation, which we consider here, and thus they have been filtered for this study by a stop word strategy.

We found that almost the entire set of 8786 natural language plans under consideration can be represented as compositions of instructions spanned up by a space of about 100 different action verbs. Among these, the most important (i.e. most frequent) actions are given by pick-and-place actions (e.g. "Place the placemat in front of the chair.") and actions for combining two or more substances (e.g. "Add the eggs to the flour."). Interestingly, the top 15 action verbs make more than 50% of all the 53,000 actions. Figure 2 shows frequencies of the top seven action verbs.

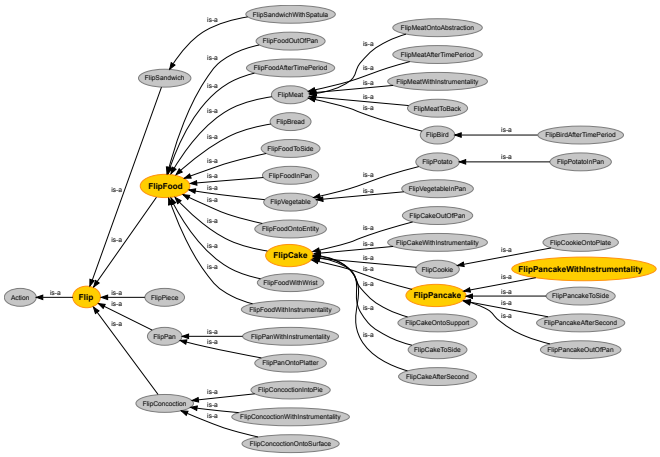


Fig. 3. Automatically generated taxonomy of differnet types of "Flipping" actions using a semantic clustering of syntactic relations.

### B. How much variation do actions exhibit?

Despite the limitation of action verbs that can be often found in the household domain with respect to numbers, the actions under consideration are still very complex. On the one hand, this is due to the world to be acted in being inherently continuous and uncertain, but, on the other hand, this also results from different actions often occurring with different parameterization in different contexts.

In order to examine how much variation the domain-relevant action verbs exhibit, we analyzed the 53,000 instructions with respect to their parameterization given by syntactic relations such as the object acted on as well as prepositional relations that modify the respective action verb, such as "with", "from", "to", "into" relations and such. These relations can have different meanings in different contexts and hence it is crucial to be able to resolve their meanings in order to understand and execute a respective action proficiently.

The dimensionality of each action verb configuration has been reduced by applying a semantic clustering technique to these parameterizations on the objects referred to in the respective relations. As a distance measure between concepts, the WuP similarity [8] has been applied to these relation arguments, whose corresponding word senses (i.e. concepts in the WordNet [9] class taxonomy) have been determined first by the importer for natural language instructions described above.

Figure 3 exemplarily shows a taxonomy of the action verb "Flip", which has been automatically generated by iteratively applying the semantic clustering algorithm to the arguments of the syntactic relations in the entire set of "flipping" instructions. It can be seen that our procedure generates a steep taxonomy of different action configurations for the action verb "Flip", which reasonably reflects different types of that action. As an example, consider the generalization of *FlippingPancakeWithInstrumentality*, which is highlighted.

Figure 4 shows the cluster sizes for relations with respect to five of the 15 most frequent action verbs. As can be seen,

| *Relation* | *Flipping* | *Cutting* | *Adding* | *Filling* | *Stirring* |
|---|---|---|---|---|---|
| *objActedOn* | 65 | 499 | 1200 | 188 | 166 |
| *prep_with* | 12 | 28 | 58 | 162 | 36 |
| *prep_on* | 2 | 34 | 42 | 3 | 4 |
| *prep_into* | 3 | 139 | 77 | 24 | 48 |
| *prep_onto* | 8 | 1 | 14 | 1 | 1 |
| *prep_to* | 2 | 26 | 460 | 24 | 18 |
| *prep_from* | 0 | 46 | 18 | 3 | 4 |
| *prep_through* | 1 | 31 | 4 | 0 | 8 |

Fig. 4. Sizes of clusters that have been obtained by semantically clustering the prepositional relations of action verbs.

a large number of diverse everyday household activities, in which an average person participates often, can be broken down to a set of elementary actions that still is very limited, though there is a wealth of different objects involved.

Our study on domain-specific action verbs supports our introductory thesis that a robot performing these activities with the same ease as humans do, needs to have a substantial body of knowledge about *how* to execute the single actions involved. With implementing robot control plans for the most important action verbs we reported on above, we expect that we can cover a wide range of everyday household activities that can be performed by our robots.

### C. Naturalistic Action Specifications

When studying the ways how humans explain complex activities to other humans, we observe that they make use of a language that is extremely underspecified and vague. As an example, consider the following sequence of instructions for making a pancake:

1) *Pour* milk into a bowl.
2) *Add* flour and *mix* well.
3) *Heat* the *greased* pan.
4) *Pour* the batter into the pan, then wait for 2 minutes.
5) *Push* the spatula under the pancake and *flip* it.
6) *Wait* for another 2 minutes.
7) *Place* the pancake on a plate.
8) *Serve*.

When formulating such directives, humans tend to omit important information, which is necessary for performing a particular action. Thus, only understanding what is explicitly *specified* by an instruction is insufficient for performing an action proficiently and successfully. The example shows that naturalistic action specifications written by humans are severely underspecified and ambiguous: instruction 2), for instance, does not specify what the flour is to be added to and neither what needs to be mixed. Similarly, the relations between the spatula, its parts and the pancake are not explicitly referred to in instruction 5).

Figure 5 depicts the network of entities and relations that specify this event more precisely. The colored entities are given by the instruction, whereas the grayed ones need to be inferred in order to obtain a deeper semantic representation. As can be seen, the information given by the natural-language directive only serves as little evidence in a complex network of dependencies between objects and actions, which needs to be completed.

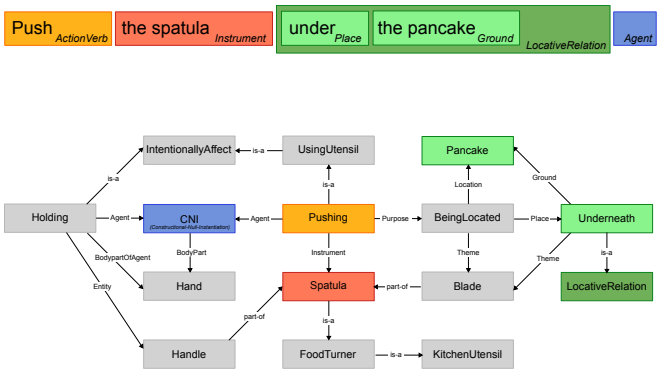


Fig. 5. Exemplary Robot Action Core for the instruction "Push the spatula under the pancake." The colored entities are given by the action specification, whereas the gray ones and the relations need to be inferred.

Typically, humans can do such inference with great ease. They share a large amount of common background knowledge about actions and objects, which is self-evident to them, rendering explicit announcements of these facts redundant and obsolete. As we will discuss in the next section, such knowledge can serve as a constraining mechanism in everyday manipulation activities as it does it in natural language. As a consequence, for completely *understanding* a naturalistic instruction with the goal of executing it, it is indispensible to not only understand what is given, but also to *infer* what is actually meant.

## III. Probabilistic Robot Action Cores

What is actually meant by an instruction specifies much more constraints than the instruction itself explicitly states. As we pointed out in the previous sections, the execution of everyday manipulation activities and human household tasks is characterized by strong experience and familiarity of an agent performing them, which makes everyday manipulation a knowledge-intensive task. Additionally, our studies on how humans explain everyday activities to others, such as cooking a meal, show that people typically omit information that is self-evident to them but inevitable for any robot that is supposed to perform them. It is the knowledge about actions, objects and the relations that hold between them that humans have in common, which allows them to infer what information is needed from what is explicitly specified.

Robots acting in human environments therefore must have corresponding action-specific knowledge available as well as mechanisms for reasoning about it, which allow to infer what to do to which objects in a particular situation. By employing such an action-specific KB, we can address the following important issues that arise when acting in human environments and interpreting naturalistic task instructions:

1) *Disambiguation*: naturalistic action specifications exhibit a high degree of ambiguity. As an example, consider the action verb "Turn over," which offers a variety of possible meanings[1]. Given contextual information, however, e.g. the objects

our background knowledge about this action substantially limits the space of possible interpretations, though there is no explicit information given about what actions to perform on which objects.
2) *Completion of actions*: naturalistic action specifications are incomplete. As pointed out above, this results from the common human knowledge about actions, which is that self-evident that no one would state it explicitly. In descriptions of activities written by humans, such as cooking recipes, we often encouter instructions like "stir occasionally". These instructions are characterized by extreme underspecification since they lack any information about the objects involved. In our example, for instance, it has to be inferred what needs to be stirred (e.g. batter), which utensil is to be used (e.g. a spoon or a mixer), where the action takes place (e.g. in a bowl) etc.

If robots are to be instructed in a natural manner, they need to be endowed with capabilities for understanding human task specifications. Previous work in this field mainly focuses on deriving goals from natural-language instructions, transforming them into a formal representation and finding an action sequence that yields the desired goal. General problem solvers have been applied to problems that require intensive mental effort and concentration, such as playing chess or solving cryptarithmetic puzzles. Everyday activities, however, are of fundamentally different nature [2]. Humans perform such activities with great ease, if not unconciously, not requiring intensive planning and deliberation. Planning about action sequences, i.e. reasoning about actions, their effects and goals, has been widely studied by both the artificial intelligence and the robotics community. However, these approaches towards action intelligence assume complete knowledge about actions and the world. As our studies show, service robots that will have to act in human environments will be faced with high uncertainty, ambiguity and underspecification, and, as Moore [10] points out, in such real-world scenarios, the knowledge about actions often is incomplete though indispensible. Only little attention has been paid to the important role that knowledge about actions plays in such environments, and about reasoning mechanisms allowing to infer the information that is needed in order to perform an activity competently.

We argue that action-specific knowledge is key to dealing with complex, human-scale everyday tasks rather than classical planning and we present a framework for modeling probabilistic dependencies between actions, objects and ontological information about the world.

The key idea of an action verb-specific knowledge base is to model events in everyday activity as generic event patterns. It has been shown that the level of abstraction of such patterns should be as abstract as possible, but as

[1]WordNet provides 11 different meanings of the verb "Turn over".

specific as necessary in order to still be able to discriminate between particular roles filled by the entities involved. As Bailey [11] points out, humans are capable of rapidly yet flexibly learning such event patterns, i.e. how different action verbs are to be used in different context. This already happens in early childhood and by hearing just a few examples. Humans are capable of abstracting away from single instances of events to more generic event patterns by building superclasses subsuming the concrete event instances. A graphical representation of such a pattern is shown in Figure 5. We are pursuing a similar strategy in building up our knowledge base.

This leads to an informal definition of an action core: a *Robot Action Core* is the set of inter- and intraconceptual *relations* that constitute an *abstract event type*, assigning an *action role* to each entity that is affected by the respective action verb. Action roles thus can be regarded as action parameters defining relations among entities involved. Knowing about all roles of a particular action in turn is required to fully specify the action under consideration.

The idea of a *Probabilistic Robot Action Core* (PRAC) is to represent a joint probability distribution over the *action roles* such that evidence given by a naturalistic instruction can be used to infer the information that is required for fully specifying the action (cmp. to Figure 5). From a probabilistic point of view, we can formulate this inference task as

$$\underset{neededRoles}{\arg\max}\ P_{Action}(neededRoles \mid givenRoles),$$

where the given roles also include relations that are incorporated by an ontology that models taxonomic (*is-a*, in symbols $\sqsubseteq$) or mereological (*part-of*, in symbols $\preceq$) knowledge. This allows us to model an action core at an appropriate level of abstraction, such that it can be applied to a large number of real-world scenarios of the same type. Given a couple of concrete single action specifications, the taxonomic relationship between entities can be exploited to abstract away to more generic action patterns. As an example, consider the following two concrete instructions:

$$\text{“Fill}_{Action}\ \text{milk}_{Theme}\ \text{into a bowl}_{Destination}\text{”}$$
$$\text{and “Fill}_{Action}\ \text{a glass}_{Destination}\ \text{with water}_{Theme}\text{”},$$

where the set $\mathcal{R} = \{Action, Theme, Destination\}$ represents the roles of the action verb “fill”. Within the PRAC framework, taxonomic generalization is used to encode

$$\text{“Fill}_{Action}\ \text{a liquid}_{Theme}\ \text{into a container}_{Destination}\text{”}.$$

as an abstract “filling” event. Given a previously unseen object, let us assume “juice”, the most probable role assignment of that word can be inferred since superclasses are taken into account in the model.

In an unknown scenario, object information can be gathered from the environment and be matched against the action core patterns, which enables to draw conclusions about what is missing and constraining the search space of possible action configurations. Here the term “scenario” is to be understood in a wide range. A scenario can be given by a real-world setting, a simulated environment, video sources or a natural language instruction.

In these terms, we can define a *Probabilistic Robot Action Core* as a conditional probability distribution $P(\mathcal{R} \times \mathcal{A} \times \mathcal{C} \mid \sqsubseteq, \preceq)$, where

| | |
|---|---|
| $\mathcal{R}$ | is the set of all action roles |
| $\mathcal{A}$ | is the set of all action verbs |
| $\mathcal{C}$ | is the set of all class concepts |
| $\sqsubseteq$ | is a taxonomy relation over $\mathcal{C}$ |
| $\preceq$ | is a mereological relation over $\mathcal{C}$. |

A Probabilistic Robot Action Core hence models a probability distribution over all possible roles of an action and can be seen as a “probabilistic typecast” on action role arguments. This probabilistic first-order representation of events can be used in order to resolve ambiguity and to complete the most plausible action specification, based on what is given by the instruction. In our examplary instruction “Flip the pancake,” the action “flip” and the object “pancake” are given, and we can use the PRAC distribution to infer action roles that are not specified, for instance the type of instrument to be used:

$$\begin{aligned}\underset{c \in Concepts}{\arg\max}\ P(\, i \sqsubseteq c \quad \mid \quad & p \sqsubseteq Pancake, \\ & Theme(a, p), \\ & ActionVerb(a, Flip) \\ & Instrument(a, i)) \\ = Spatula & \end{aligned} \tag{1}$$

In the next section, we show how such a PRAC model can be designed and learned from labeled language data.

## IV. Implementation

Here we present our implementation of PRAC, which is applicable to natural-language instructions. Figure 6 shows the overall architecture of our system. We mainly use five knowledge sources which are publicly available in order to construct the relational PRAC model. In particular, we use

1) the *WordNet* lexical database for conceptual and taxonomic knowledge,
2) the *FrameNet* database for action and role definitions,
3) the *Stanford Parser* for extracting syntactic information from natural-language instructions,
4) the `wikihow.com` website for domain-specific action knowledge,
5) the *Amazon Mechanical Turk* marketplace in order to obtain semantically labeled ground truth data.

The first three knowledge sources are depicted in the upper part of the diagram in Figure 6. Knowledge from each of these resources is relational in its nature, such that it can be directly be imported into a logic framework like PRAC. The lower part shows knowledge sources that we use for the purpose of data acquisition, which is discussed in the next section.

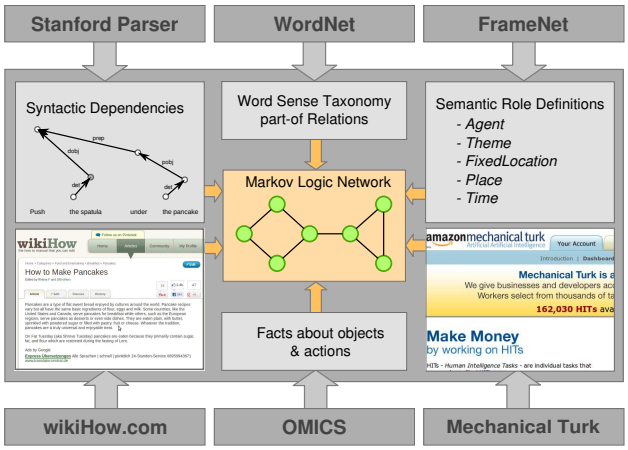


Fig. 6. Implementation architecture of the PRAC Framework.

### A. Knowledge Sources

*WordNet* [9] is a lexical database that groups words of equal conceptual meanings into groups, the so-called *synsets*. Additionally, it provides a deep taxonomy of these synsets as well as mereological relations. Given a word and its part of speech, the possible word meanings can be obtained from WordNet, which corresponds to the sets $\mathcal{C}, \mathcal{A}, \sqsubseteq$ and $\preceq$ in the PRAC model. *FrameNet* [12] provides conceptualizations of actions that consist of an action definition and a set of associated roles that represent the parameters of a respective action. An action itself is represented as an abstract concept and specific action verbs represent instances of these concepts. As an example, consider the defintion of the action concept *MovingInPlace*, which is defined in FrameNet as follows:

> *A* Theme *moves with respect to a* fixed location*, generally with a certain* Periodicity*, without undergoing unbounded translational motion or significant alteration of configuration/shape.*

Possible instances of this abstract event are given by the action verbs *Rotate*, *Shake*, *Spin*, *Twirl*, *Flip*, *Turn around* etc., which share the same action parameters, such as the *Theme* undergoing a non-translational motion, the *FixedLocation* or the *Periodicity* of the motion. Figure 7 shows definitions of a selection of roles attached to the *MovingInPlace* event.

Beside semantic knowledge about actions, objects and their interrelations, also syntactic information can be strong evidence in understanding naturalistic action specifications. Prepositional relations, for instance, but also the sequence of words in an instruction can have strong influence on its semantics. For example, consider the instruction "Flip the pancake with a spatula", where the prepositional relation *with* indicates an instrumental relationship between the action *flip* and the subsequent word *spatula*. Other prepositional relations, such as "into", "onto", "from" or "off", however, indicate *Goal* and *Source* relations in instructions like "Fill a cooking pot *with* water *from* the tap."

Information given by such syntactic relations can be incorporated by employing natural-language parsing. In particular, syntactic dependencies connect constituents in a natural-language instruction by means of binary predicates that can be directly imported into a relational model such as PRAC. The following dependencies, for instance, are obtained by the Stanford Parser [13], being applied to the instruction above:

1. *det*(*pancake-3*, *the-2*)
2. *dobj*(*Flip-1*, *pancake-3*)
3. *det*(*spatula-6*, *a-5*)
4. *prep_with*(*Flip-1*, *spatula-6*)

These dependencies indicate that the second word "the" depends on the third word "pancake" as a determiner, "pancake" represents a direct object of the verb "Flip" and "spatula" depends on "Flip" via a prepositional "with" relation. However, PRAC does not rely on a correct parse, for it considers these relations just as evidence features. In connection with the semantic action roles defined by FrameNet, the syntactic Stanford Dependencies represent the set $\mathcal{R}$ of interconceptual relations.

| ***Role*** | ***Definition*** |
|---|---|
| *Theme* | A physical *entity* that is participating in non-translational motion. |
| *FixedLocation* | The point or set of points that define the limits of motion for the *Theme*. |
| *Angle* | The amount of rotation that the *Theme* undergoes |
| *Periodicity* | The number of times the *Theme* returns to a state in a given duration. |
| *Direction* | The direction of rotation of the *Theme*. |
| *Place* | The location where the bounded motion happens. |
| *Time* | The time at which the Theme is in bounded motion. |

Fig. 7. Role Definitions for the Action Concept *MovingInPlace*

### B. Data Sources

The lower part of Figure 6 represents three Web resources that we use for acquiring data. We extracted more than 1,400 natural-language instructions from the "Food & Entertaining" category of `wikihow.com` and had them semantically annotated. This set of instructions comprises the action verbs *add*, *cut*, *fill*, *flip*, *mix*, *place*, *pour* and *put*, which belong to the most frequent actions verbs we reported on in Section II-A. In a first step, we employed Amazon Mechanical Turk to crowdsource the task of semantically annotating word senses for each of the nouns, verbs, adjectives and adverbs occurring in the natural-language instructions. The possible word senses have been obtained from the WordNet lexical data base. In a second step, we annotated the action roles in each instruction. As a result, we obtained a semantically annotated corpus of natural language data that can be used in order to train the PRAC models described above. In the near future, we also plan to incorporate data from the *Open Mind Indoor Commonsense* (OMICS) [14] data base, which has not yet been integrated in our current implementation.

### C. Markov Logic Networks

We implemented the PRAC knowledge base as a *Markov Logic Network* (MLN) [15]. MLNs represent a knowledge

| *Predicates* | |
|---|---|
| *hasSense*$(w, r!, i)$ | *hasRole*$(w, r!, i)$ |
| *isa*$(s, s)$ | *hasPOS*$(w, pos!, i)$ |
| *det*$(w, d, i)$ | *dobj*$(w, w, i)$ |
| *prep_from*$(w, w, i)$ | *prep_with*$(w, w, i)$ |
| *prep_into*$(w, w, i)$ | *conj_and*$(w, w, i)$ |

| # | *Formula* |
|---|---|
| 1 | *hasSense*$(w_1, +s_1, i) \wedge$ *hasRole*$(w_1, +r_1, i) \wedge$ *isa*$(+s_1, +s_2)$ |
| 2 | *hasSense*$(w_1, +s_1, i) \wedge$ *hasRole*$(w_1, +r_1, i) \wedge$ *hasSense*$(w_2, +s_2, i) \wedge$ *hasRole*$(w_2, +r_2, i)$ |
| 3 | *prep_from*$(w_1, w_2, i) \wedge$ *hasRole*$(w_1, +r_1, i) \wedge$ *hasRole*$(w_2, +r_2, i)$ |
| 4 | *prep_with*$(w_1, w_2, i) \wedge$ *hasRole*$(w_1, +r_1, i) \wedge$ *hasRole*$(w_2, +r_2, i)$ |
| 5 | *prep_into*$(w_1, w_2, i) \wedge$ *hasRole*$(w_1, +r_1, i) \wedge$ *hasRole*$(w_2, +r_2, i)$ |
| 6 | *conj_and*$(w_1, w_2, i) \wedge$ *hasRole*$(w_1, +r, i) \wedge$ *hasRole*$(w_2, +r, i)$ |
| 7 | *hasPOS*$(w, +pos, i) \wedge$ *hasRole*$(w_1, +r, i)$ |

Fig. 8. Predicate declarations and formulas representing the PRAC model for the action *fill*.

representation formalism that combines first order logic and probabilistic undirected graphical models (*Markov Random Fields*). An MLN can be regarded as a first-order knowledge base with a weight attached to each formula. The probability of a possible world $x$ is defined as

$$P(X = x) = \frac{1}{Z} \exp\left(\sum_i w_i n_i(x)\right),$$

where $n_i(x)$ is the number of true groundings of the $i$-th formula $F_i$ in the world $x$, $w_i$ is the weight attached to $F_i$ and $Z$ is a normalizing constant.

It is known that learning and inference in MLNs are computationally very expensive. Hence, in order to keep the implementation tractable, we simplify the PRAC distribution to factorize according to

$$\prod_{A \in \mathcal{A}} P\left(\mathcal{R} \times \mathcal{C} \mid A, \sqsubseteq, \preceq\right),$$

i.e. we have one PRAC model for each action verb. Figure 8 shows the predicates and formulas that have been chosen for the MLN. In the predicate declarations, the "!" operator specifies a functional constraint on the respective predicate argument, i.e. all entities fill exactly one action role in one particular action specification (as indicated by the *hasRole* predicate declaration).

In an MLN, the logical formulas specify features of the data at hand and represent dependencies among relations. Here, the "+" operator indicates that the respective formula will be expanded with respect to the corresponding argument domain. The first formula, for instance, specifies a correlation between the conceptual meaning of a particular word $w_1$ in an instruction $i$, its action role $r_1$ and its super classes in the class hierarchy. This formula therefore models the generalization principle from concrete event instances to more generic event patterns. The second formula assumes correlations between co-occurences of action roles and their types, and Formulas 3-7 model the connection between syntactic and semantic features.

### D. Experiments

Due to the computational complexity of learning and inference in MLNs, we had to restrict our experiments to a small excerpt of the WordNet class taxonomy as well as a limited number of action-specific relations and formulas in the current implementation. However, this small subset suffices for demonstrating the strength of the PRAC concept.

As a first proof of concept, we conducted experiments on a very small number of training instructions (i.e. 3) and analyzed the generalization performance of the model on a larger set of test instructions (i.e. 10). Here, the objects involved in the test set have not been part of any training instruction. The training instructions were given by "Fill a cup with water from the tap," "Fill milk into a bowl," and "Fill the cup and the pot with water," where we used the action roles *Theme*, *Source* and *Goal*. After having learned the PRAC model for the action verb "fill" with only three instances, the system succeeded to correctly assign the corresponding word meanings and action roles of 10 out of 10 natural language directives, each of which addressed objects that have not been part of the training set. Examples are "Fill a glass with juice" or "Pour oil into the pan." Here, evidence was given only by the syntactic features obtained from the parser.

Inferring information that is missing in an instruction can be done in a two step process: First, the instruction is analyzed for the given action roles as just described. Afterwards, the action roles assigned to the words in the instruction are taken as evidence, and for each missing action role, a new entity is introduced, assigned to the respective role (according to Eq. 1). Given the instruction "Fill the sink with water," PRAC is then able to determine the action roles given by the instruction and to infer e.g. where to get the water from (i.e. *Source=Tap*).

Although being conducted using very small data sets, the results of the experiments show that PRAC succeeds to automatically find general event patterns out of a set of concrete action instantiations, which can be applied to new situations with previously unseen objects. Using PRAC, knowledge about actions can be acquired not only by examining concrete instructions from natural language plans, but also more general statements about objects and their action-specific roles can be incorporated in PRAC learning. A rule such as "Always fill liquids into containers" can be specified and, due to the class hierarchy in PRAC, effects will be propagated to all subclasses of *Liquid*.

## V. Related Work

In recent years, much work has been done in order to make knowledge sources available to robots, which are indented for human use [7], [16], and to generate robot plans out of natural-language instructions [7], [17]–[20]. Dzifcak *et al.* [19] use a combinatorial categorial grammar for deriving a goal formulation in temporal logics in order to find an action sequence that achieves this goal. Matuszek *et al.* [17] use statistical machine translation techniques to match natural-language navigation directives

against a formal path description language. Others [16], [18] use probabilistic models to derive plans to be executed by a robot. What all these approaches have in common is that they do not take into account that natural-language instructions typically are severely underspecified and ambiguous. They make what is commonly referred to as the *closed world assumption* postulating that all knowledge about the world is given and complete. For a very limited set of actions, as in robot navigation, for example, this seems reasonable. When tasks become more complex, however, such as preparing a meal, this is an uncommon form of human cognition [2], [21]. Additionally, most approaches to teach robots by means of natural language are designed to capture and execute what is specified by an instruction using "shallow" mappings to robot control, but they are not intended to accumulate more general action knowledge that can be recalled in different situations.

Our work goes beyond those approaches by modeling abstract event patterns in a probabilistic first order knowledge base, taking prior ontological knowledge about actions and objects into account. This allows to abstract away from concrete action instances towards a more generic notion of actions that helps in disambiguating and completing underspecified naturalistic action specifications. Our work is not about finding action sequences given a particular goal, but about *how* to perform complex everyday activities in presence of partial and incomplete information.

## VI. Conclusions

In this work we argue that robots that are to perform complex everyday activities must be equipped with a substantial body of action-specific knowledge in order to resolve the problems of ambiguity and underspecification, which are ubiquitous when dealing with naturalistic action descriptions. We reported on a domain analysis, which supports this thesis. We introduced the concept of Probabilistic Robot Action Cores, which are a novel, knowledge-driven approach to model abstract event types, and we presented promising experimental results that show that these models are well-suited to be learned from natural-language data and show excellent generalization performance. Future investigations will concentrate on finding more efficient algorithms for probabilistic first-order reasoning in order to build more expressive PRAC models and to evaluate them on larger data sets.

We believe that equipping robots with action-specific knowledge is a key paradigm for implementing more flexible, cognitive robot behavior and pushing autonomous robots to performing more advanced everyday activities.

## VII. Acknowledgements

This work has been supported by the EU FP7 Project RoboHow[2] (grant number 288533) and the CoTeSys cluster of excellence (Cognition for Technical Systems[3]), part of the Excellence Initiative of the German Research Foundation (DFG).

[2] http://www.robohow.eu

[3] http://www.cotesys.org

## References

[1] M. Beetz, U. Klank, I. Kresse, A. Maldonado, L. Mösenlechner, D. Pangercic, T. Rühr, and M. Tenorth, "Robotic Roommates Making Pancakes," in *11th IEEE-RAS International Conference on Humanoid Robots*, Bled, Slovenia, October, 26–28 2011.

[2] J. Anderson, "Constraint-directed improvisation for everyday activities," Ph.D. dissertation, Dissertation, 1995.

[3] H. Simon, "Rational choice and the structure of the environment." *Psychological review*, vol. 63, no. 2, p. 129, 1956.

[4] L. Kunze, M. E. Dolha, and M. Beetz, "Logic Programming with Simulation-based Temporal Projection for Everyday Robot Object Manipulation," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, San Francisco, CA, USA, September, 25–30 2011, best Student Paper Finalist.

[5] L. Kunze, M. Tenorth, and M. Beetz, "Putting People's Common Sense into Knowledge Bases of Household Robots," in *33rd Annual German Conference on Artificial Intelligence (KI 2010)*. Karlsruhe, Germany: Springer, September 21-24 2010, pp. 151–159.

[6] D. Jain, L. Mösenlechner, and M. Beetz, "Equipping Robot Control Programs with First-Order Probabilistic Reasoning Capabilities," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2009, pp. 3626–3631.

[7] M. Tenorth, D. Nyga, and M. Beetz, "Understanding and Executing Instructions for Everyday Manipulation Tasks from the World Wide Web," in *IEEE International Conference on Robotics and Automation (ICRA)*, Anchorage, AK, USA, May 3–8 2010, pp. 1486–1491.

[8] Z. Wu and M. S. Palmer, "Verb semantics and lexical selection," in *ACL*, 1994, pp. 133–138.

[9] C. Fellbaum, *WordNet: an electronic lexical database*. MIT Press USA, 1998.

[10] R. Moore, "A formal theory of knowledge and action," DTIC Document, Tech. Rep., 1984.

[11] D. Bailey, "When push comes to shove: A computational model of the role of motor control in the acquisition of action verbs," Ph.D. dissertation, UNIVERSITY of CALIFORNIA, 1997.

[12] C. Fillmore, "Frame semantics and the nature of language," *Annals of the New York Academy of Sciences*, vol. 280, no. 1, pp. 20–32, 1976.

[13] M. De Marneffe, B. MacCartney, and C. Manning, "Generating typed dependency parses from phrase structure parses," in *Proceedings of LREC*, vol. 6, 2006, pp. 449–454.

[14] Honda Research Institute USA Inc., "Open Mind Indoor Common Sense," http://openmind.hri-us.com, 2008.

[15] P. Domingos and M. Richardson, "Markov logic: A unifying framework for statistical relational learning," in *Proceedings of the ICML-2004 Workshop on Statistical Relational Learning and its Connection to Other Fields*. Banff, Canada: IMLS, 2004, pp. 49–54.

[16] J. Ryu, Y. Jung, K. Kim, and S. Myaeng, "Automatic extraction of human activity knowledge from method-describing web articles," *Proceedings of the 1st Workshop on Automated Knowledge Base Construction*, p. 16, 2010.

[17] C. Matuszek, D. Fox, and K. Koscher, "Following directions using statistical machine translation," in *Proceeding of the 5th ACM/IEEE international conference on Human-robot interaction*. ACM, 2010, pp. 251–258.

[18] S. Tellex, T. Kollar, S. Dickerson, M. Walter, A. Banerjee, S. Teller, and N. Roy, "Understanding natural language commands for robotic navigation and mobile manipulation," in *Proc. NatâĂŹl Conf. on Artificial Intelligence (AAAI)*, 2011.

[19] J. Dzifcak, M. Scheutz, C. Baral, and P. Schermerhorn, "What to do and how to do it: Translating natural language directives into temporal and dynamic logic representation for goal management and action execution," in *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*. IEEE, 2009, pp. 4163–4168.

[20] E. Neo, T. Sakaguchi, and K. Yokoi, "A natural language instruction system for humanoid robots integrating situated speech recognition, visual recognition and on-line whole-body motion generation," in *Advanced Intelligent Mechatronics, 2008. AIM 2008. IEEE/ASME International Conference on*. IEEE, 2008, pp. 1176–1182.

[21] R. Barker, *Ecological psychology: Concepts and methods for studying the environment of human behavior*. Stanford Univ Pr, 1968.