2009 IEEE International Conference on Robotics and Automation
Kobe International Conference Center
Kobe, Japan, May 12-17, 2009

Affordance based word-to-meaning association

V. Krunic G. Salvi A. Bernardino L. Montesano J. Santos-Victor

Abstract— **This paper presents a method to associate meanings to words in manipulation tasks. We base our model on an affordance network, i.e., a mapping between robot actions, robot perceptions and the perceived effects of these actions upon objects. We extend the affordance model to incorporate words. Using verbal descriptions of a task, the model uses temporal co-occurrence to create links between speech utterances and the involved objects, actions and effects. We show that the robot is able form useful word-to-meaning associations, even without considering grammatical structure in the learning process and in the presence of recognition errors. These word-to-meaning associations are embedded in the robot's own understanding of its actions. Thus they can be directly used to instruct the robot to perform tasks and also allow to incorporate context in the speech recognition task.**

I. Introduction

To interact with humans, a robot needs to communicate with people and understand their needs and intentions. The by far most natural way for a human to communicate is language. This paper deals with the acquisition by a robot of language capabilities linked to manipulation tasks. Our approach draws inspiration from infant cross situational word learning theories that suggest that infant learning is an iterative bootstrapping process [12]. It occurs in an incremental way (from simple words to more complex structures) and involves multiple tasks such as word segmentation, speech production, and meaning discovery. Furthermore, it is highly coupled with other learning process such as manipulation, for instance, in mother infant interaction schemes [8].

Out of the multiple aspects of language acquisition, this paper focuses on the ability to discover the meaning of words through human-robot interaction. We adopt a developmental robotics approach [18], [10] to tackle the language acquisition problem. In particular, we consider the developmental framework of [11] where the robot first explores its sensory-motor capabilities. Then, it interacts with objects and learns their affordances, i.e. relations between actions and effects. The affordance model uses a Bayesian network to capture the statistical dependencies among a set of robot basic manipulation actions (e.g. grasp or tap), object features and the observed effects by means of statistical learning techniques exploiting the co-occurrence of stimuli in the sensory patterns.

The main contribution of the paper is the inclusion in the affordance model [11] of verbal descriptions, provided by a human, of the robot activities. The model exploits temporal co-occurrence to associate speech segments to the meanings in terms of actions, object properties and the corresponding effects. Despite we do not use any social cues or the number and order of words, the model provides the robot with the means to learn and refine the meaning of words in such a way that it will develop a rough understanding of speech based on its own experience.

Our model has been evaluated using a humanoid torso able to perform simple manipulation tasks and to recognize words from a basic dictionary. We show that simply measuring the frequencies of words with respect to a self-constructed model of the world, the affordance network, is enough to provide information about the meaning of these utterances even without considering prior semantic knowledge or grammatical analysis. By embedding the learning into the robot's own task representation, it is possible to derive links between words such as nouns, verbs and adjectives and the properties of the objects, actions and effects. We also show how the model can be directly used to instruct the robot and to provide contextual information to the speech recognition system.

The rest of the paper is organized as follows. After discussing related work, Section III briefly describes, through our particular robotic setup, the problem and the general approach to be taken in the learning and exploitation phases of the word-concept association problem. Section IV presents the language and manipulation task model and the algorithms used to learn and make inferences. In Section V we describe the experiments and provide some details on the speech recognition methods employed. Results are presented in Section VI and finally, in Section VII, we conclude our work and present ideas for future developments.

II. Related Work

Computational models for cross situational word learning have only been studied recently. One of the earliest works is the one by Siskind [15] who proposes a mathematical model and algorithms for solving an approximation of the lexical-acquisition task faced by children. The paper includes computational experiments, using a rule based logical inference system, that shows that the acquisition of word-to-meaning mappings can be performed by constraining the possible meanings of words given their context of use. They show that acquisition of word-to-meaning mappings might be possible without knowledge of syntax, word order or reference to properties of internal representations other than

V. Krunic, G. Salvi, A. Bernardino, L. Montesano, J. Santos-Victor are with the Instituto de Sistemas e Robótica, Instituto Superior Técnico, Lisboa, Portugal. {vkrunic,gsalvi,alex,lmontesano,jasv}@isr.ist.utl.pt

G. Salvi is currently with the Speech, Music and Hearing lab at the Royal Institute of Technology (KTH), Stockholm, Sweden

This work was supported by EU NEST Project 5010 - Contact, and by Fundação para a Ciência e Tecnologia (ISR/IST plurianual funding) through the POS Conhecimento Program that includes FEDER funds.

978-1-4244-2789-5/09/$25.00 ©2009 IEEE

co-occurrence. This has motivated a series of other research in cross-situational learning.

Frank, Goodman and Tenenbaum [5] presented a Bayesian model for cross-situational word-learning that learns a ”word-meaning” lexicon relating objects to words. Their model explicitly deals with the fact that some words do not represent any object, e.g., a verb or an article. By modeling the speaker’s intentions, they are also able to incorporate social cues typically used by humans. Dominey and Voegtlin [4] propose a system extracting meaning from narrated video events. The system requires the knowledge of the grammatical construction of the narrations. Some recent works have also studied robot language acquisition based on self-organizing neural networks [6] or word-object associations through incremental one class learning algorithms [9]. Focusing on object names, [17] exploits object behavior (resulting effects of an action) to create object categories using reinforcement learning.

Probably, the closest work to ours is presented in [19], [20], where a human subject was instrumented with devices to perceive its motor actions, speech discourse and the interacting objects (camera, data glove and microphone), and an automatic learning system was developed to associate phoneme sequences to the performed actions (verbs) and observed objects (nouns). Common phoneme patterns were discovered in the speech sequence by using an algorithm based on Dynamic Programming. These patterns were then clustered into similar groups using and Agglomerative Clustering Algorithm in order to define word-like symbols to associate to concepts.

In our case, the robot has access to its own action which eases the word-meaning association. Also, we assume that the robot has already learned a set of words and is able to recover them from the acoustic signal. We leave out of the current study the problem of learning the words from sequences of acoustic classes as in [19], [16] and learning the acoustic classes from the speech signal as in [14]. In spite of these simplifying assumptions, in this study, objects are represented by their features (shape, color, size) rather than by their category, thus allowing for a more flexible description. As a result, our model automatically incorporates adjectives (object properties). We also deal with learning the description of the effects (outcomes of actions), therefore addressing the acquisition of concepts related to behaviors (e.g “the ball is moving”, “the box is still”).

III. Approach

In this section, we provide an overview of the full system. As mentioned before, we assume that the robot is at a developmental stage where basic manipulation skills have already been learned up to a maturity level that includes a model of the results of this actions on the environment (see [11] for further details). In order to make the presentation less abstract, we describe the particular robotic setup used in the experiments and the skills already present in the system.

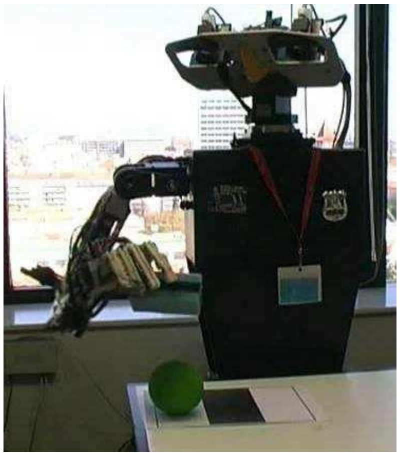

Fig. 1. Baltazar, the humanoid torso used in the experiments.

A. Robot skills and developmental stage

We used Baltazar, a 14 degrees of freedom humanoid torso composed by a binocular head and an arm (see Figure 1).

The robot is equipped with the skills required to perform a set of simple manipulation actions denoted by a_i on a number of objects. In our particular experiments we consider the actions grasp, tap and touch. In addition to this, its perception system allows it to detect objects placed in front of it and extract information about them. More precisely, it extracts simple visual features that are clustered in an unsupervised way to form symbolic descriptions of object characteristics such as color, size or shape. We denote with f_1, f_2 and f_3 the color, shape and size descriptor labels of objects. When performing an action, the robot can also detect and categorize the effects produced by its actions. Effects are mainly identified as changes in the perception such as the object velocity (e_1), the velocity of the robot’s own hand (e_2) and the persistent activation of the contact sensors in the hand (e_3).

Based on this action-perception basic skills, the robot has also undergone a training period that allowed it to establish relations between actions, object features and effects[1]. This model captures the world behavior under the robot actions. It is important to note that the model includes the notion of consequences[2] and, up to a certain extent, an implicit narrative structure of the execution of an action upon an object.

The robot is also equipped with audio perception capabilities that allow it to recover an uncertain list of words based on a previously trained speech recognizer.

B. Incorporating speech

Given this state of the robot, we aim to exploit the co-occurrence of verbal descriptions and simple manipulation tasks to associate meanings and words. Our approach is the following. During the execution of an action, the robot

[1] This is not strictly necessary in the model presented in the next section. However, in order to test the expressiveness of the method we made this assumption.

[2] One should be always careful about causality inference. However, under certain constraints one can at least guess about induced statistical dependencies [13].

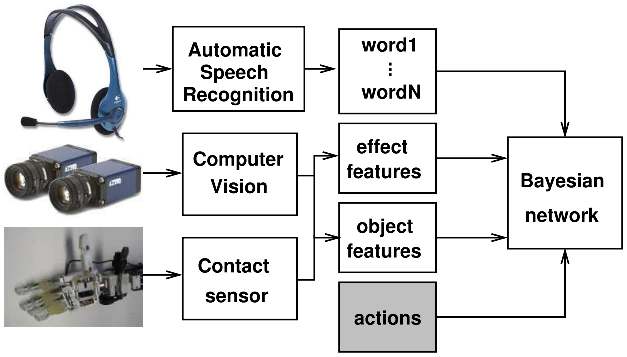

Fig. 2. Overview of the setup.

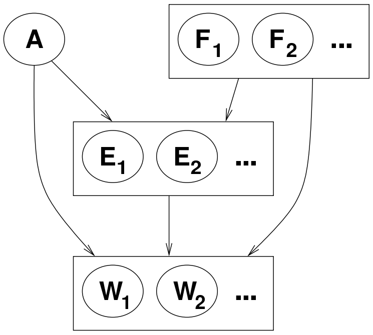

Fig. 3. Graphical representation of the model. The affordance network is represented by three different sets of variables: actions (A), object features (F_i) and effects (E_i). Each word w_i may depend on any subset of A, F_i and E_i.

listens to the users speech and recognizes some words of the speech stream and stores them in a bag of words ($\{w_i\}$), i.e. an unordered set where multiple occurrences are merged. These words are correlated with the concepts of actions, object features and effects present in the world. Our objective is to learn the correct relationships between the word descriptions and the previous manipulation model through a series of robot-human interaction experiments. These relations implicitly encode word-meaning associations grounded to the robot's own experience.

We will model this problem in a Bayesian probabilistic framework where the actions A, defined over the set $\mathcal{A} = \{a_i\}$, object properties F, over $\mathcal{F} = \{f_i\}$ and effects E, over $\mathcal{E} = \{e_i\}$ are random variables. We will denote $X = \{A, F, E\}$ the state of the world as observed by the robot. The joint probability $p(X)$ encodes the basic world behavior grounded by the robot through interaction with the environment. The verbal descriptions are denoted by the set of words $W = \{w_i\}$. Figure 2 illustrates all the information fed to the learning algorithm.

If we consider the world concepts or meanings being encoded by X, then, to learn the relationships between words and concepts, we estimate the joint probability distribution $p(X, W)$ of actions, object features, effects, and words in the speech sequence. Once good estimates of this function are obtained, we can use it for many purposes, for example:

- to compute associations between words and concepts, by estimating the structure of the joint pdf $p(W, X)$;
- to plan the robot actions given verbal instructions from the user in a given context, through $p(A, F \mid W)$;
- to provide context to the speech recognizer by computing $p(W \mid X)$.

In the following section we detail the methods and techniques used to learn and exploit word-meaning associations.

IV. Model - Algorithms

In this section, we present the model and methods used to learn the relations between words and the robot own understanding of the world. Our starting point is the affordance model presented in [11]. This model uses a discrete Bayesian network to encode the relations between the actions, object features and the resulting effects. The robot learns the network from self-experimentation with the environment and the resulting model captures the statistical dependencies among actions, object features and the consequences of the actions.

In this paper, we extend the previous model to include also information about the words describing a given experience. Recall that X denotes the set of (discrete) variables representing the affordance network. For each word in W, let w_i represent a binary random variable. A value $w_i = 1$ indicates the presence of this word, while $w_i = 0$ indicates the absence of this word in the description. We impose the following factorization over the joint distribution on X and W

$$P(X, W) \quad = \quad \prod_{w_i \in W} p(w_i \mid X_{w_i}) p(X). \tag{1}$$

where X_{w_i} is the subset of nodes of X that are parents of word w_i. The model implies that the set of words describing a particular experience depends on the experience itself[3]. On the other hand, the probability of the affordance network is independent of the words and, therefore, is equal to the one in [11]. Figure 3 illustrates our model.

In our model, each variable w_i is a discrete random variable that indicates the presence or not of a word according to the particular configuration of the affordance network. A strong assumption of our model is the independence among words. This is actually known as the *bag of words* assumption and is widely used, for instance, in document classification ([2]), and information retrieval. Given a network structure, i.e. the set of X_{w_i} per each word w_i, our model simply computes the frequency of such a word for each configuration of the parents.

We are also interested in selecting among all the possible models described by Eq. 1 that best fit the data. This model selection problem has been widely studied in the machine learning literature (see [7] for a review). In our case, we use a variation of the simple greedy approach known as K2 algorithm [3] to select the most likely graph given a set of data

$$G^* = \arg max_g p(D \mid G) \tag{2}$$

where $D = \{(X_i, W_i)\}$ represents the training data, i.e. a set of pairs of world meanings and verbal descriptions.

[3] This point requires a careful treatment when dealing with baby language learning and, usually, explicit attention methods are required to constraint the relations between words and the meanings they refer to.

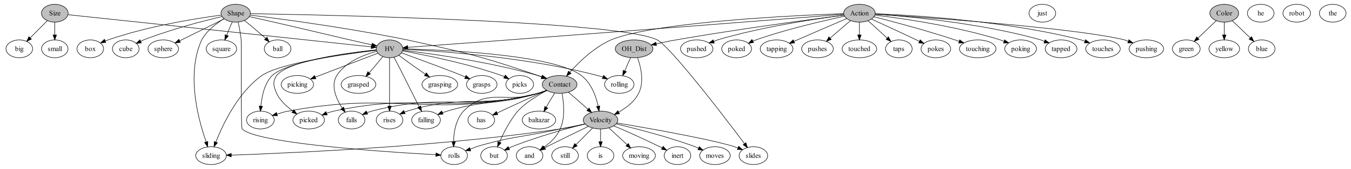

Fig. 4. Graph of the full Bayesian network

Note that despite the fact we may have a huge number of nodes, our model restricts the set of possible networks to the factorization in Eq. 1. As a result the space to search is reduced considerably. However, we may loose some dependencies among words that are part of speech.

Finally, let us briefly describe some inference queries that can be solved by our model once learned. As mentioned in Section III, the network allows to perform several speech based robot-human interactions. First, the robot can be easily instructed to perform a task. This corresponds to recovering the (set of) action(s) given the words W_s provided by the recognizer, e.g. $p(A \mid W_s)$. When dealing with a particular context, i.e. a set of potential objects to interact with, the robot may maximize:

$$
\begin{aligned}
< a^*, o^* > &= \arg max_{a_i, o_i \in O_s} p(a_i, F_{o_i} \mid W_S) \quad (3)\\
&\propto \prod_{w_i \in W_s} p(w_i \mid a_i, F_{o_i}) p(a_i, F_{o_i}) \quad (4)
\end{aligned}
$$

where O_s is the set of objects detected by the robot and F_{o_i} the features associated to object o_i. Assuming that we have non informative priors over the actions and objects, the robot seeks to select the action and object pair that maximizes the probability of W_s. Alternatively, the robot may compute the k-best pairs.

The proposed model also allows to use context to improve recognition. Consider the case where the recognizer provides a list of possible sets of words. The robot can perform the same operation as before to decide what set of words is the most probable or rank them according to their posterior probabilities. In other words, one can combine the confidence of the recognizer on each sentence with the context information to select.

V. Experiments

In this section we describe the experimental protocol taken in the word-concept learning phase.

A. Verbal Description of the Experiences

Each experience from [11] was verbally described by a number of observers with utterances in a predefined form. Each utterance describes first the action the robot performs on a certain object and then the effects that the action has produced. Examples of this are: “Baltazar is grasping the ball but the ball is still.”, “The robot touches the yellow box and the box is moving.”, “He taps the green square and the square is sliding.”. Each action, object property and effect is represented by a varying number of synonyms for a total of 49 words. Three alternative descriptions of each experience were considered.

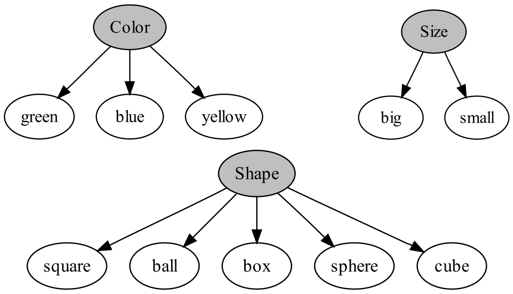

Fig. 5. Object properties words

B. Speech input

In order for the robot to learn from the verbal descriptions, we equipped it with hearing capabilities. As discussed in Section I, we assume that one of the basic skills of the robot is the ability to classify speech input into sequences of words.

The speech-to-text unit is implemented as a hidden Markov model (HMM) automatic speech recognizer (ASR). Each word belonging to the language described above is modeled as an HMM with a number of states proportional to the phonemic length of the word. Additionally a three-state-model is defined in order to model silence.

A set of recordings were used to train the model parameters. These include single words recordings for model initialization and utterances in the form described above for training. The recordings were performed by 17 non-native speakers of English. The hardware and location of the recordings was freely chosen by the speakers and usually involved a computer and headsets with a close microphone. Only orthographic transcriptions were available with no time stamps and the ASR models were trained with the Baum-Welch iterative algorithm [1].

No grammatical structure other than a simple loop of words was imposed to the decoder at run time, in agreement with our hypothesis that a grammar is not necessary in order to learn simple word-meaning associations. The performance of the recognizer was computed with a leave-one-out technique by iteratively training the models on all but one speaker in the data set and testing on the remaining speaker. The resulting percentage of correctly classified words was about 81%.

VI. Results

We first describe in this Section the word-meaning associations acquired by the robot and, then, exemplify the possible use of this model.

A. Learning

The results of learning word meaning associations are displayed in Figure 4 and detailed in the following figures.

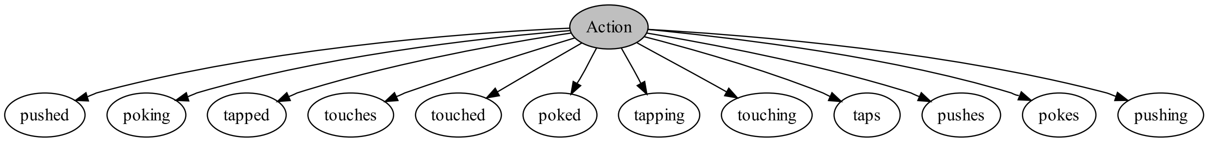

Fig. 6. Action words (excluding grasping)

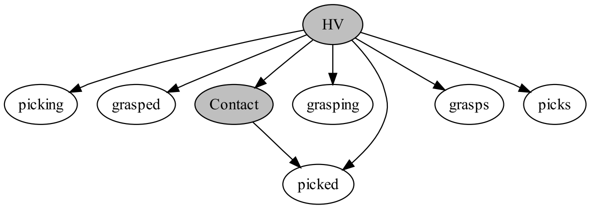

Fig. 7. Action words (grasping)

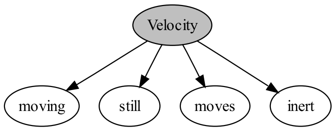

Fig. 8. Effect words: generic movement

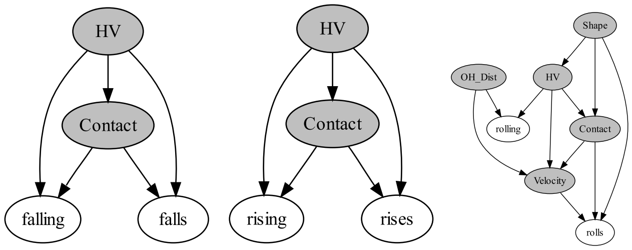

Fig. 9. Effect words: some specific movements

Figure 4 displays the full graph of the Bayesian network, where the affordance nodes are filled whereas word nodes have white background. The full network is included to give an impression of the overall complexity of the model. In the following, we will focus on subsets of words, in order to simplify the discussion.

Some of the word nodes do not display any relationship with the affordance nodes. The so called *non referential* words are: "robot", "just", "the", "he". This result is not surprising if we notice that the affordance network did not include a representation of the robot itself ("robot", "he"), nor a representation of time ("just"). Moreover, articles and auxiliary verbs were also expected to be non referential.

Words expressing *object features* are displayed in Figure 5. These are clearly linked to the right affordance node. This result is in accordance with previous research that showed that it is possible to learn word object associations.

Words expressing *actions* are displayed in Figure 6 and 7. In general (Figure 6) action words are correctly linked to the Action node in the affordances. For words indicating the specific action *grasp*, the link is to the node HV (hand velocity after contact), and in one case to Contact[4]. The reason for this is that, in our data, HV is high only for grasping actions. The information on hand velocity is, therefore, sufficient to determine whether a grasp was performed. Moreover, HV can only assume two values (high, low), while Action can assume three values (grasp, tap and touch), thus making the first a more concise representation of the concept grasp.

Words describing *effects* usually involve more affordance nodes. In case of words indicating generic movement the link is to the object velocity node, as expected (see Figure 8). In case of more specific movements the results are shown in Figure 9. The figure shows how the words for rising and falling are connected to nodes that can describe successful and unsuccessful grasps. In fact, rising is only observed in our data in case of a successful grasp and falling in case of an unsuccessful grasp. Hand velocity (HV), as explained earlier, is a compact representation of the action grasp. The presence or absence of hand-object contact, on the other hand, determines if the grasping was successful. This is an example where a more complex concept is created by combining more than one affordance node. One example in which the interpretation of the model is harder is given by the words for rolling in the same figure (which includes the hand velocity and the relative hand object velocity (OH_DIST)).

B. Using the model

Some possible uses of the word meaning association model are described in this Section.

Table I shows some examples of using incomplete verbal descriptions to assign a task to the robot. The robot has a number of objects in its sensory field (represented by the object features in the first column in the Table). The Table shows, for each verbal input W_S (column) and each set of object features F_{o_i} (row), the best action computed by Equation 3 when the set of objects O_s is restricted to a specific object o_i. The global maximum over all actions and objects for a given verbal input, corresponding to the general form of Equation 3, is indicated in bold face in the table.

If the combination of object features and verbal input is incompatible with any actions, $P(a_i, F_{o_i} \mid W_S)$ may be 0 $\forall a_i \in \mathcal{A}$. These cases are displayed with a dash in the Table. In case this happens for all available objects (as for "ball sliding" in the example), the behavior of the robot is not defined. A way to solve such cases may be, e.g., to initiate an interaction with the human in order to clarify his/her intentions.

Another application of our model is to use the knowledge stored in the Bayesian network to disambiguate between possible interpretations of the same speech utterance, given

[4] When we repeated the experiment using the true transcriptions of the utterances, the link disappeared, suggesting that it was due to recognition errors. Note, however, that Contact is compatible with the action pick.

TABLE I

EXAMPLES OF USING THE BAYESIAN NETWORK TO SELECT ACTIONS AND OBJECTS

				Verbal input			
objects on the table	"small grasped"	"moving green"	"ball sliding"	"big rolling"	"has rising"	"sliding small"	"rises yellow"
light green circle big	-	grasp, p=0.034	-	**tap, p=0.227**	grasp, p=0.019	-	-
yellow circle medium	-	-	-	-	**grasp, p=0.073**	-	**grasp, p=0.3**
dark green box small	grasp, p=0.122	grasp, p=0.041	-	-	grasp, p=0.037	**tap, p=0.25**	-
blue box medium	-	-	-	-	grasp, p=0.037	-	-
blue box big	-	-	-	tap, p=0.022	grasp, p=0.017	-	-
dark green circle small	**grasp, p=0.127**	**tap, p=0.127**	-	-	grasp, p=0.064	-	-

TABLE II

EXAMPLES OF USING THE BAYESIAN NETWORK TO IMPROVE ASR

	N-best list from ASR (N=3)		
objects on the table	"tapping ball sliding"	**"tapping box slides"**	"tapped ball rolls"
light green circle big	0	0	0.0567
yellow circle medium	0	0	0.0567
dark green box small	0	0.0605	0
blue box medium	0	0.0589	0
blue box big	0	0.0605	0
dark green circle small	0	0	0.0567

the context. The speech recognizer can return an N-best list of hypotheses, ranked by the acoustic likelihood. Our model provides a natural way of revising such ranking by incorporating information of the situation the robot is currently facing.

An example is shown in Table II. Similarly to Table I, Table II shows a situation in which a number of objects are in the range of the robot's sensory inputs. The utterances corresponding to each column in the Table are, this time, the simulated output of a speech recognizer in the form of an N-best list with length three. The other difference from Table I is that the probabilities in each entry are computed by integrating over all possible actions.

The second hypothesis, in bold face in the Table, is selected when the posterior probability over all possible actions and objects is computed. This in spite of the fact that the recognizer assigned a best match to the hypothesis in the first column.

VII. CONCLUSIONS AND FUTURE WORK

This paper proposes a common framework to model affordances and to associate words to their meaning. The model exploits co-occurrence between its own actions and a description provided by a human to infer the correct associations between words and actions, object properties and action's outcomes. Experimental results show that the robot is able to learn clear word-to-meaning association graphs from a set of 49 words and a dozen of concepts with just a few hundred human-robot-world interaction experiences. The learnt associations were then used to instruct the robot and to include context information in the speech recognizer. Based on these results, there are many extensions for our language acquisition model, specially to relax some of our assumptions. In particular, we plan to include more complex robot-human interaction and social cues to allow a less rigid language between the instructor and the robot.

REFERENCES

[1] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist*, 41(1):164–171, 1970.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res*, 3:993–1022, 2003.

[3] G.F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347, 1992.

[4] Peter Ford Dominey and Thomas Voegtlin. Learning word meaning and grammatical constructions from narrated video events. In *Proceedings of the HLT-NAACL 2003 workshop on Learning word meaning from non-linguistic data*, pages 38–45, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

[5] M. Frank, N. Goodman, and J. Tanenbaum. A bayesian framework for cross-situational word learning. *Advances in Neural Information Processing Systems*, 20, 2008.

[6] Xiaoyuan He, Tomotaka Ogura, Akihiro Satou, and Osamu Hasegawa. Developmental word acquisition and grammar learning by humanoid robots through a self-organizing incremental neural network. *IEEE Transactions on Systems, Man, and Cybernetics–Part B*, 37(5), 2008.

[7] D. Heckerman. A tutorial on learning with bayesian networks. In *In M. Jordan, editor, Learning in graphical models*. MIT Press, 1998.

[8] F. Lacerda, E. Marklund, L. Lagerkvist, L. Gustavsson, E. Klintfors, and U. Sundberg. On the linguistic implications of context-bound adult-infant interactions. In *Proceedings of the Fourth International Workshop on Epigenetic Robotics*, 2004.

[9] L. Lopes and L. Seabra. How many words can my robot learn?: An approach and experiments with one-class learning. *Interaction Studies*, 8(1), 2007.

[10] M. Lungarella, G. Metta, R. Pfeifer, and G. Sandini. Developmental robotics: a survey. *Connection Science*, 15(40):151–190, December 2003.

[11] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor. Learning object affordances: From sensory motor maps to imitation. *IEEE Transactions on Robotics, Special Issue on Bio-Robotics*, 24(1), 2008.

[12] L. Montague N. Akhtar. Early lexical acquisition: the role of cross-situational learning. *First Language*, 19(57):337–358, 1999.

[13] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2000.

[14] G. Salvi. Ecological language acquisition via incremental model-based clustering. In *Proc. of Eurospeech/Interspeech*, pages 1181–1184, Lisbon, Portugal, 2005.

[15] Jeffrey Mark Siskind. A computational stdy of cross-situational techniques for learningword-to-meaning mapping. *Cognition*, 61:39–91, 1996.

[16] Veronique Stouten, Kris Demuynck, and Hugo Van hamme. Discovering phone patterns in spoken utterances by non-negative matrix factorisation. *IEEE Signal Processing Letters*, 15:131–134, 2008.

[17] S. Takamuku, Y. Takahashi, and M. Asada. Lexicon acquisition based on object-oriented behavior learning. *Advanced Robotics*, 20(10):1127–1145, 2006.

[18] Juyang Weng. The developmental approach to intelligent robots. In *AAAI Spring Symposium Series, Integrating Robotic Research: Taking The Next Leap*, Stanford, USA, Mar 1998.

[19] Chen Yu and Dana H. Ballard. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perception*, 1(1):57–80, 2004.

[20] Chen Yu and Dana H. Ballard. A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, 70(13-15):2149–2165, 2007.