Associating word descriptions to learned manipulation task models

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. This knowledge is acquired by the robot in an unsupervised way by self-interaction with the environment. When a human user is involved in the process and describes a particular task, the robot can form associations between the (co-occurrence of) speech utterances and the involved objects, actions and effects. We extend the affordance model to incorporate a simple description of speech as a set of words. We show that, across many experiences, 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. Word-to-meaning associations are then used to instruct the robot to perform tasks and also allow to incorporate context in the speech recognition task.**

I. Introduction

Nowadays, robots are required to work in social environments (hospitals, museums, homes) and to interact with humans. Learning and adaptation have emerged as a programming paradigm to cope with the highly dynamic, unstructured and stochastic scenarios where the the robots operate. When interacting with humans, a robot also needs to communicate with people to 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. This is a first step towards the long term objective of developing robots able to learn language through interaction with humans.

Our approach draws inspiration from infant cross situational word learning theories that suggest that infant learning is an iterative bootstrapping process [14]. This learning procedure is highly complex and starts very early in the first months of life, following a developmental paradigm. The acquisition of language capabilities requires the continuous interaction between the learner and the teacher. It also 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].

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

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.

According to the previous discussion, we adopt a developmental robotics approach [18], [11] to tackle the language acquisition problem. In particular, we consider the developmental framework of [13]. Initially, the robot explores the capabilities of its own body, both in term of motor actions and perception, and learns to execute and perceive basic actions. When it masters (to a certain point) the control of its own body, it starts interacting with objects and learns the effects of their actions upon the objects (affordances). After having acquired good enough models of how objects behave, the robot is ready to interact with humans, using the object affordance model as a link to perceive and imitate others' actions, as well as to predict their intentions. This can be seen as an implicit form of communication, whereby a human teacher can describe a task to the learner by means of showing many examples of the execution task [10].

In this paper we focus, out of the multiple aspects of language acquisition, on the ability to link previously learned models of manipulation (as the affordance model described above) to verbal descriptions provided by a human. At the moment, these descriptions are formed by a set of words provided by a trained speech recognizer. We use the affordance Bayesian network model of [13]. Roughly speaking, the network captures 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.

We extend the previous model to consider the utterances spoken by the user as input data, and use the same learning mechanisms to associate speech segments to the meanings – actions, object properties and effects. Currently, we do not use any social cues, nor the number and order of words. The objective is to provide the robot with the means to learn and refine the meaning of words in such a way that it will develop an understanding of speech based on its own experience. This type of learning, exploiting the redundancy and co-occurrence of stimuli in multiple situations, is called "cross-situational" and will allow the development of more natural human-robot-interaction interfaces.

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 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.

A. Related Work

Computational models for cross situational word learning have only been studied recently. Perhaps one of the earliest works is the one by Siskind [17] 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 co-occurrence. This has motivated a series of other research in cross-situational learning.

Yu and Ballard [20] developed a model based on machine translation methods and time co-occurrence of the utterances and visual information about the objects or actions. They also include non-speech contextual information such as the speaker's gaze direction, head direction, hand movements which allows incorporating extra information about the actual intentions of the speaker. 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].

Probably, the closest work to ours is presented in [19], 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. This association was done computing the probability of each word given every possible meaning and then running the EM algorithm to compute the maximum likelihood association.

The proposed work differs from [19] in the following aspects:

- Whereas in [19] the information about the performed action is hidden in the perceptual data, in our case this information is explicit. Once the robot has decided to perform a certain action, the action value is deterministic.
- The work in [19] deals explicitly with the object concept. In this work, following the affordances model described in [13], objects are represented by their features (shape, color, size) rather by their category, thus allowing a more flexible description of objects, using not only their nouns, but also their properties (adjectives).
- Our work also deals with learning the description of the effects (outcomes of actions), therefore addressing the acquisition of concepts related to the behaviors (e.g "the ball is moving", "the box is still").
- The work in [19] focused only on the learning problem. In this work we want to address also the use of speech to instruct the robot to perform tasks. The task description and the mode of execution will be determined by concepts transmitted verbally to the robot.

This rest of the paper is organized as follows. In Section II we briefly describe, 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 III presents the language and manipulation task model and the algorithms used to learn and make inferences. In Section IV we describe the experiments and provide some details on the speech recognition methods employed. Results are presented in Section V and finally, in Section VI, we conclude our work and present ideas for future developments.

II. 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 [13] 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.

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 II).

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

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

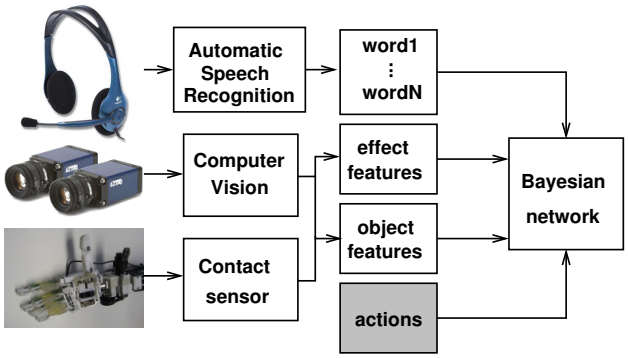

Fig. 2. Overview of the setup.

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 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 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 the 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.

III. 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 [13]. This model uses a discrete Bayesian network to encode the relations between the actions, object features and the resulting effects. The robot is able to learn the network from self-experimentation with the environment and the resulting model captures the statistical dependencies among action, 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 denote 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

[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 [15].

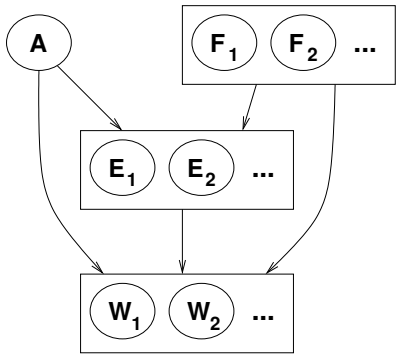

Fig. 3. Graphical representation of the model. The affordance network is represented by three different sets of variables: actions (A), object features (F_i) and effects (E_i). Each word w_i may depend on any subset of A, F_i and E_i.

factorization over the joint distribution on X and W

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

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, thus the affordance part of the model is equal to the one in [13]. 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

$$G^* = \arg max_g p(D \mid G) \quad (2)$$

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

[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.

Finally, let us briefly describe some inference queries that can be solved by our model once learned. As mentioned in Section II, 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.

$$< a^*, o^* > = \arg max_{a_i, o_i \in O_s} p(a_i, F_{o_i} \mid W_S) \quad (3)$$

$$\propto \prod_{w_i \in W_s} p(w_i \mid a_i, F_{o_i}) p(a_i, F_{o_i}) \quad (4)$$

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.

IV. 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 [12] 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.

B. Speech input

In order for the robot to learn from the verbal descriptions, we equipped it with hearing capabilities. We assume that one of the basic skills of the robot is the ability to classify speech input into sequences of words. This assumption was motivated by the focus of the current work on associating words and meanings. A thoroughly developmental approach to speech classification would involve learning the words from sequences of acoustic classes as in [19] and, even, learning the acoustic classes themselves from the data as in [16]. We leave these developments for future work.

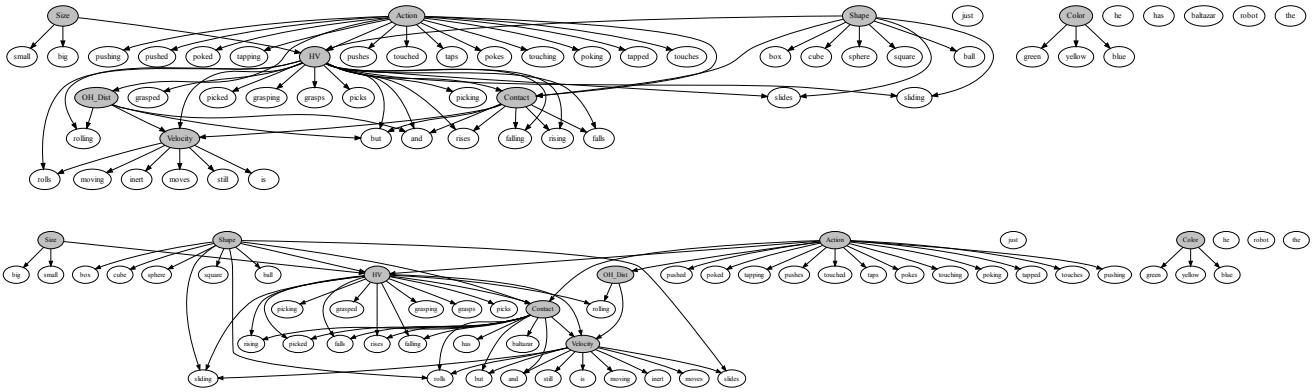

Fig. 4. Full network, top: synthetic data, bottom: recognizer

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%. The decoder produces, besides confusions between words, also a number of insertions and deletions of words in the utterance. In these experiments, the relative number of insertions and deletions was not controlled, resulting in roughly double as many insertions than deletions. Short words such as “and”, “but”, “the”, “he” and “is” are likely to be inserted. The word “the” was also deleted in many utterances.

V. 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.

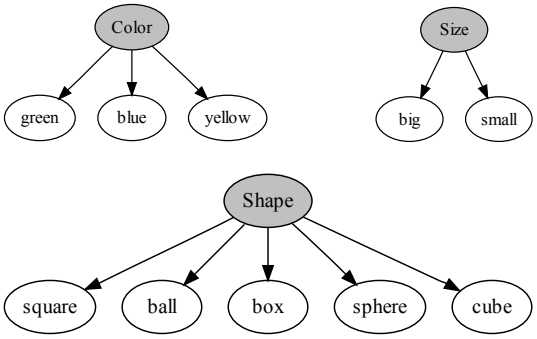

Fig. 5. Object properties words

Figure 4 displays two graphs corresponding to the Bayesian networks learned with perfect speech recognition (top), and with the real speech recognizer (bottom). In each graph, affordance nodes are filled whereas word nodes have white background. The full networks were included to give an impression of the overall complexity of the model and to observe that the errors from the recognizer had a small effect in learning the word meaning associations.

In the following we will always refer to the results obtained with real speech recognition and, in order to simplify the discussion, we will focus on subsets of words.

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” (plus “Baltazar” and “has” in the perfect recognition case). This result is not surprising if we notice that the affordance network did not include a representation of the robot itself (“Baltazar”, “robot”, “he”), nor a representation of time (“just”). Moreover, articles and auxiliary verbs were also expected to be non referential. The exception to this is the verb “is” that is linked to the node Velocity. This requires further explanation, but is probably due to asymmetries in the data set, i.e. an unbalanced set of examples from our strict verbal descriptions.

Words expressing *object features* are displayed in Figure 5.

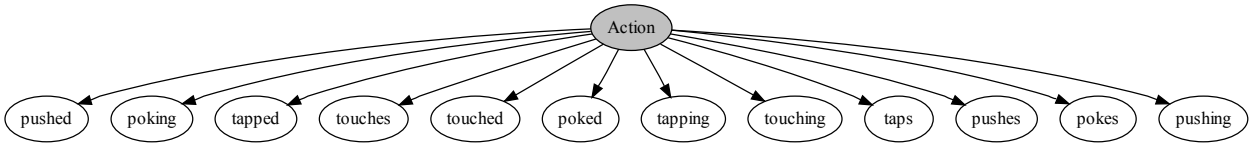

Fig. 6. Actions words (excluding grasping)

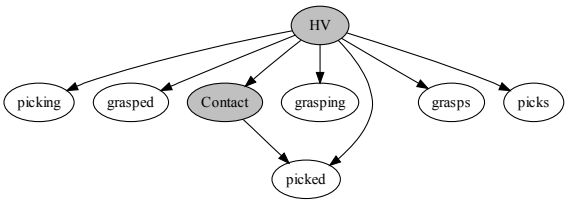

Fig. 7. Action words (grasping)

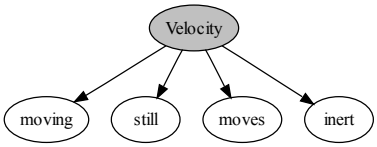

Fig. 8. Effect words: generic movement

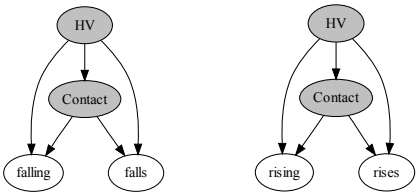

Fig. 9. Effect words: some specific movement

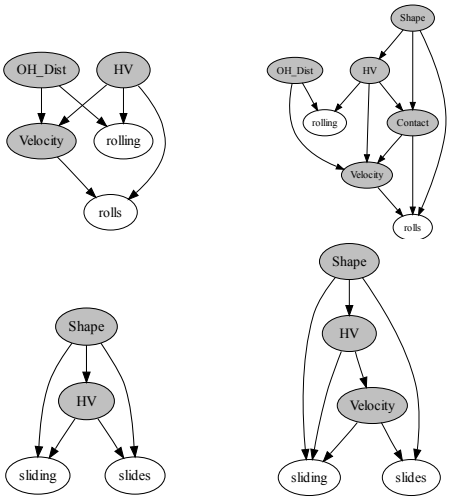

Fig. 10. Effect of the errors in the speech recognizer: perfect recognition (left) and real recognition (right)

These are clearly linked to the right affordance node. This result is in accordance with previous research that showed that it was 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), and in one case to Contact, even though the latter was not present with perfect recognition. 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 movement the results are shown in Figure 9 and Figure 10 where the effects of recognition errors are also illustrated. Figure 9 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.

Comparing the results we obtained with perfect recognition and the real speech recognizer, we observe that most of the dependencies inferred by the model were the same. Two examples in Figure 10 illustrate the cases were we observed a difference in the Bayesian network. The figure depicts two alternative graphs for the words “rolls”, “rolling” (top) and the words “slides”, “sliding” (bottom). The left graphs are obtained with perfect speech recognition, whereas the right graphs with the real recognizer. In general the recognition errors add extra complexity to the data. This results in the introduction of new dependencies in the model. The extra nodes that are linked to the words are, however, related to the concepts the words stand for. For example, in the case of “sliding”, Velocity was added, which is consistent with

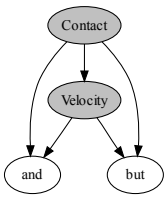

Fig. 11. Conjunctions "and" and "but"

movement. In the case of "rolling", the inclusion of Shape is related to the fact that only balls can roll, whereas the inclusion of Contact probably and artifact and should be explained by looking at the recognition results in details.

Finally the conjunctions "and" and "but" are linked to the affordance nodes Contact and Velocity as depicted in Figure 11. There is no simple explanation of this association, and we believe the reason for this is two-fold. Firstly this result is strongly affected by recognition errors that mostly involve short words, as pointed out in the previous Section. If we consider the perfect recognition case, the conjunctions are linked to three nodes: object-hand velocity, hand velocity and contact. By inspecting the probabilities of the word for each combination of the values of these nodes, we can infer that "and" is always used in conjunction with a successful action, while "but" is associated with unsuccessful grasp. This is not surprising because the action that is most likely to fail in our data is grasp. More in general, associations that involve different concepts in the experiments, as in this case, are simply more difficult to learn, than direct associations as, e.g., with object properties. Dealing with more complex concepts probably requires a richer set of experiences, and therefore a larger and more complete data set.

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 the context. The Viterbi decoder in the speech recognizer can be configured, e.g., to return an N-best list of hypotheses for the transcription of a given utterance, the entries in the list being ranked with the corresponding likelihood. This is used when building dialog systems because the correct transcription may not be the first in the list, and the dialog manager may use context from the interaction with the user to select the hypothesis that is most consistent with the situation.

TABLE II

EXAMPLES OF USING THE BAYESIAN NETWORK TO SELECT MULTIPLE INTERPRETATIONS OF THE INPUT UTTERANCE

	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

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 higher probability to the hypothesis in the first column. This was, however, a preliminary result based on a simulation. In order to prove the effectiveness of the method for this application, experiments with the real speech recognizer should be used.

VI. CONCLUSIONS AND FUTURE WORK

A. Conclusions

We have shown how a robot can learn the meaning of words in manipulation tasks by exploring the correlations between speaker's utterances and the sensory information related to actions, objects and outcomes of its own experimentation of the world. The learning process does not use the order of the words on the speech signal, thus making it more flexible toward changes in speaking style, it is known, e.g., how spontaneous speech may fail to comply to the grammatical structure of the specific language. It also facilitates the usage of the acquired knowledge in instructing the robot to perform tasks. In this case the formulation of a task may assume a different form than the description of a manipulation experiment.

Experimental results show that the robot is able to learn clear word-to-meaning association graphs from a set of 49

TABLE I

EXAMPLES OF USING THE BAYESIAN NETWORK TO SELECT ACTIONS AND OBJECTS

objects on the table	Verbal input "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	-	-

words and a dozen of concepts with just a few hundred human-robot-world interaction experiences.

B. Future Work

Despite the encouraging results, the proposed model is a first step towards a more complete model that will allow to capture the language acquisition process more accurately. In particular we seek to relax some of the main assumptions done in this paper.

The first assumption is the existence of a predefined set of words and the ability of the robot to extract the identity of such words from an acoustic input. This allows us to learn all the word-meaning associations at once. Working directly with the audio signal, or at a phonemic level, is a more realistic setup for our developmental learning approach. We are currently studying the way to incorporate word discovery in our model. This will require to develop incremental strategies and inclusion of social cues as in [5], [20] to cope with the huge search spaces of the full problem. In particular, we are considering bootstrapping our model with a few words, e.g. by word spotting, and then incrementally adding words by associating the unknown part of the speech input with different nodes in the affordance network. Another important direction for research is to look into the interaction mechanisms that allow children to develop language.

The second assumption is that the language used to instruct the robot is rigidly defined and does not resemble the natural interaction between parents and their children. This has allowed us to learn from a limited number of experiences because each utterance contains an almost complete description of the situation. The drawback is that any asymmetry in the data may result in spurious dependencies in the model as noted in Section V. The introduction of an incremental approach, as described above, may allow us to relax this assumption and cope with very sparse descriptions of each experience by focusing at each time step on a limited set of possible word-meaning associations. This will hopefully increase the robustness of the method to asymmetries in the data.

Finally we assume no interactions between words. By linking words to the affordance network, we allow some dependencies among them. However language has a rich structure where word order, grammar and other context situations play an important role. By relaxing this assumption we might be able to model part of this complexity and even to build more abstract associations where phrases, instead of single words, are linked to meanings.

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] Manuel Lopes, Francisco S. Melo, and Luis Montesano. Affordance-based imitation learning in robots. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2007.

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

[12] L. Montesano, M. Lopes, A. Bernardino, and J. Santos-Victor. Affordances, development and imitation. In *IEEE - International Conference on Development and Learning*, 2007.

[13] 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.

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

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

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

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

[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.