A Probabilistic Approach to Learning a Visually Grounded Language Model through Human-Robot Interaction

Haris Dindo and Daniele Zambuto

Abstract— **Language is among the most fascinating and complex cognitive activities that develops rapidly since the early months of infants' life. The aim of the present work is to provide a humanoid robot with cognitive, perceptual and motor skills fundamental for the acquisition of a rudimentary form of language. We present a novel probabilistic model, inspired by the findings in cognitive sciences, able to associate spoken words with their perceptually grounded meanings. The main focus is set on acquiring the meaning of various perceptual categories (e.g. red, blue, circle, above, etc.), rather than specific world entities (e.g. an apple, a toy, etc.). Our probabilistic model is based on a variant of multi-instance learning technique, and it enables a robotic platform to learn grounded meanings of adjective/noun terms. The systems could be used to understand and generate appropriate natural language descriptions of real objects in a scene, and it has been successfully tested on the NAO humanoid robotic platform.**

I. INTRODUCTION

We have investigated the lexical acquisition problem, particularly how a robot can be bootstrapped into communication and what are the necessary prerequisites for robots in order to learn a language. This work focuses on three of the earliest problems that robots need to solve as they acquire their native language: (a) identifying the meaning of words grounded in perceptual data, (b) associating these meanings with lexical units, and (c) inferring a rudimentary grammar for further understanding and interaction. Lexical acquisition seems to be innately driven by the principle of reference: words refer to objects, actions, and attributes of the environment. The robot must acquire the possible meanings of words from their non-linguistic (perceptual) input, and determine which co-occurrences are relevant from a multitude of potential co-occurrences between words and entities in the environment while acquiring syntactic rules that encodes word order and phrase structure constraints. Observational learning may be used to deduce word meanings from cross-situational experiences. Joint attention plays an important role in learning terms of reference. Infants are more likely to connect words with their referents when engaged in joint attention with their caregivers [3] and have certain biases which constrain the set of possible meanings of words [6][1]. We have used these assumptions to bootstrap the lexical acquisition process. Concepts acquired from lexical acquisition can be used to initialize a logic representation of several observable entity of world. Language acquisition therefore proceed in parallel with concept acquisition. Concepts underlying acquired language model can be considered as independent from language acquisition process and can be reused for other cognitive tasks. The ultimate goal of the proposed system is to take advantage of acquired concepts, and language model to engage in simple dialogue with a human partner. We have eliminated some simplifying constraints present in many related works, and improved the performance and robustness of the algorithm by using robust probabilistic techniques.

The rest of this paper is organized as follows. The next section briefly describes previous approaches to language acquisition and symbol grounding problem. In section III we describe our lexical acquisition algorithm. Finally, we present experimental results of the model on a robotic platform, and outline conclusions and future works.

II. RELATED WORKS

There has been a huge interest in grounded language acquisition in the past years. In literature there are numerous examples of language acquisition systems inspired by different theories and implemented with different methodologies, ranging from hard-coded systems to neural networks and probabilistic learning systems. In this section we describe some of the most interesting systems, which have in part influenced our work.

Visual Translator system [9] (VITRA) is a natural language generation system which is grounded directly in perceptual input. From a sequence of digitized video frames low-level sensory processes perform recognition and tracking of visible objects. Detailed domain knowledge is used to categorize spatial relations between objects, and dynamic events. Higher level propositions are formed from these representations which are mapped to natural language using a rule-based text planner. In contrast to other works, VITRA is not designed as a learning system.

[13] applied the principle of grounding words semantics in sensory inputs of robots to acquisition and evolution of artificial language. These experiments are computational model of language evolution, based on a naming name, and therefore can not be used for interactions with human agents, but only artificial agents. [4] applied neural networks to symbol grounding problem by connecting sensorimotor inputs with arbitrary symbolic representations via category-invariance detectors. The system allow to learn single words phrase referred to simple image. [12] approached the problem of acquisition of natural categories and labels by robots from the point of view of perceptual grounding. CELL (Cross-channel Early Lexical Learning) [12] is a system able to

Authors are with the Department of Computer Science, University of Palermo, Palermo, Italy (dindo,zambuto@dinfo.unipa.it)

learn object names from a corpus of spontaneous infant-directed speech, and to process single and two-word phrases which referred to the color and shape of objects. CELL seems to be the first model of language acquisition which learns words and their semantics from raw sensory input without any human-assisted preparation of data. Associations between linguistic and contextual channels are chosen on the basis of maximum mutual information. Another system which learns from raw sensory data was presented in [8]. The authors focused on the challenges arising from the headset-free learning of speech labels and natural interactive learning. In particular they present a mechanism for auditory attention integrating bottom-up and top-down information for the segmentation of the acoustic stream. In DESCRIBER [11] is addressed the problem syntactic structure acquisition within a grounded learning framework. Learning algorithms acquire probabilistic structures which encode the visual semantics of phrase structure, word classes, and individual words. Using these structures, a planning algorithm integrates syntactic, semantic, and contextual constraints to generate natural and unambiguous descriptions of objects in novel scenes.
Another related approach is TWIG [7], a word learning system that allows a robot to learn compositional meanings for new words that are grounded in its sensors. TWIG allows a robot (1) to learn the meanings of deictic pronouns, (2) to contrast new word definitions with existing ones, thereby creating more complex definition, and (3) to use words learned in an unsupervised manner for production, comprehension, or referent inference. The techniques that TWIG introduces are extension inference and word definition tree. Its technique are more generally applicable to other word categories, including verbs, prepositions and nouns.

Our work while not making significant advances compared to the systems presented, puts more emphasis on one fundamental problem in language acquisition process: the search for the referent. We endow the system with a real model of attention and formalize a multi-instance learning algorithm for the acquisition of semantic categories. Our goal is to create robust learning algorithms, able to build knowledge even in absence of important pragmatic information.

III. MODEL OVERVIEW

In our previous work we focused on the learning of grounded language models from examples [5]. In the experiment we proposed, the demonstrator could chose one of the objects of the scene and provide its, more or less detailed, description. In that case, the referent of a descriptive phrase was directly given to the robot. This information, while on one hand simplifies the learning process and allows the robot to discard the majority of the incorrect associations, on the other reduces the applicability of the technique to more complex environments and does not allow any level of interaction with the demonstrator. In this work, we wanted to relax this constraint by making the interaction between the demonstrator and a robot more natural in the teaching phase.

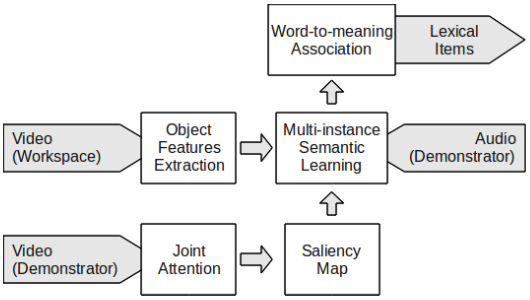


Fig. 1. General model overview.

Without knowing the referent of the sentence (e.g. the object being described by the demonstrator), the language learning problem becomes more difficult. Indeed, the robot should maintain a huge number of assumptions, many of which are incorrect, that would make the association of possible meanings to available words computationally impossible. In the new experimental setting, the teacher first tries to capture the attention of the robot by fixing his gaze on the object of interest or points it, and once obtained the attention of the robot, she describes the object. The robot is unable to determine with certainty the object of interest, but she may exploit the direction of the gaze and the pointing direction of the demonstrator to filter the possible referents, without other a priori given knowledge.

We equipped the robot with some basic skills to simulate the process of joint attention. The robot is able to recognize the demonstrator (face, hand) and to detect any activity (speech, hand movements, etc.). Moreover, it is able to estimate the direction of the gaze and recognize the pointing actions. These information is merged in order to determine the salience of each object in the scene. The robot first tries to determine the area to the maximum salience on the work surface, by following gaze direction and hand gestures, then observes and stores the objects that correspond to this area. The robot can also indicate the area of interest for further feedback from the demonstrator. When the demonstrator says something (presumably the object description), the robot stores the most salient objects and the transcript of the statement that he heard, which will constitute the training set for learning. The sample will be discarded when the degree of salience is not high enough, and the robot was unable to identify with certainty the most salient objects. Using a humanoid robot platform, the demonstrator can guess the state of the robot by using the same mechanism of joint attention and correct it, if necessary, so to minimize the ambiguity in the training set. Before describing the problems addressed in the present work we provide operational definitions of several terms which are used throughout this article. A *semantic category* (or semantic unit) specifies a range of sensory inputs which can be grouped and associated with a word/symbol. For example, a semantic category might specify a portion of the color spectrum. Such a semantic category could be used to ground the semantics for a color term such as “red”. A *semantic class* specifies a set of semantic categories grounded in the same sensory channel. For example, a semantic class could be used to associate acquired color terms (color class). A *lexical item* encodes the association between a word and its corresponding semantic category.

All semantic categories are derived from visual sensory signal. Feature extractors computes visual features from the sensors (video). Each extracted feature encodes relevant, non-redundant, information from the visual sensory stream about observable proprieties of the world (semantic class). Potential visual features include categories of shape, color, size, and spatial relation (see Table I). Any word may potentially be paired with any semantic category which is derived from the same utterance-context pair. These pairs are clustered to generate a set of lexical items.

Each sample collected composed of a bag of words provided in the description, the sensory characteristics of multiple objects and a level of salience associated with each object. The salience can be regarded as an a priori estimate of the possible referent of the statement. Imagine that the robot should determine the degree of associability between the word red and the color of the object. If the estimated salience is correct in most cases, each example containing the word red contain at least one instance of the color red (RGB feature), but it will also contain instances of different colors that are an additional source of ambiguity. In order to correctly infer the meaning of the word red, the robot must be able to isolate from each example the proper instance, discarding all others. This problem in literature is known as *multi-instance learning*. In multi-instances learning the labels are only assigned to bags of instances (i.e. labels are not assigned to individual instances). In the binary case, a bag is labeled positive if at least one instance in that bag is positive, and the bag is labeled negative if all the instances in it are negative.

In our previous work [5], a semantic distortion metric was used to select appropriate semantic category from several hypothetical ones. The meaning of each word (i.e. it's semantic category) was treated as a random variable and modeled with a multivariate Gaussian distributions. These distributions are estimated for each semantic class (shape, color, size). Taking the example above, the algorithm estimates a semantic category (Gaussian density) for each sensory channel (shape, color, size) from the set of positive examples associated with the word red. Each of these categories represents a hypothesis about the possible meaning of the word, and a hypothetical association between the semantic class and word. Similarly, the algorithm estimates a probabilistic model from the negative examples associated with the word red, i.e. those examples where the word is not present. The semantic grounding is done with the semantic class (and the associated semantic category) that maximizes the semantic distortion measure between the two probabilistic models. The previous algorithm then consists of two basic steps: (a) the estimation of semantic categories and their negative models (background probability) for each acquired word and (b) the association of meaning-word obtained by probabilistic measures on the estimated probabilities.

In this work, we have maintained the same structure as the previous algorithm. Again, we first estimate the semantic categories and negative models and then use probabilistic methods to determine the correct association. The first difficulty, as already mentioned, is precisely in the estimation of densities: treating them with bags of instances and not with individual instances complicates the learning problem, which becomes multi-instance problem. The estimation process must take into account a priori information obtained from the attention system (salience), and at the same time find the set of redundant instances for each class.

We present a new algorithm for learning semantic categories, inspired by some multi-instance learning techniques [10]. In this work, we also present a new algorithm for semantic association that, compared to the previous work, also integrates information related to the learned syntax, as well as those related to sensory observations alone (semantics). Words that belong to the same semantic class, must follow the same syntactic rules, and thus should belong to the same syntactic class. The system is schematically outlined in Fig. 1.

A. Semantic clustering as multi instances learning

The first phase of the algorithm deals with the estimation of semantic categories and negative models. For each word w recognized by the system, there is a set of training data, consisting of positive and negative examples, i.e. examples where the word is used or not used. Each sample consists of a number of instances and the degree of salience associated with them. For example, the word red, assuming we have three semantic classes, get three sets of positive examples, and three sets of negative examples, one for each class. We denote positive bags as $\mathbf{x}_n$, and the ith instance in that bag as $\mathbf{x}_{ni}$. Suppose each instance can be represented by a real-valued feature vector. Likewise, $\mathbf{x}_n^-$ denotes a negative bag and $\mathbf{x}_{ni}^-$ is the ith instance in that bag. For each semantic class-word pair, we then estimate two probability density: the distribution of feature values conditioned on the presence of word $p(\mathbf{x}|c,w)$ (hypothetical semantic category) and the distribution of feature values conditioned on the absence of word $p(\mathbf{x}|c,\overline{w})$ (background distribution).

We want to estimate a parametric probability density $p(\mathbf{x}|c,\overline{w},\boldsymbol{\theta}^-)$ from all negative bags within each class. Unlike the classical paradigm of multi-instance learning, we can not be sure that the bag contain only negative instances of the concept to be learned. For example, if we want estimate the negative model of the word red, we can use examples that describe green objects, but we can not be sure that the bags do not also contain instances of red object (that is, the process of attention may have estimated a high degree of salience for one red object next to the object described). We must carefully select the bags to be used for estimation of the negative model. A simple procedure to minimize this type of error is to select examples where the degree of salience is concentrated on few objects only. We assume that the data points $\mathbf{x}_{ni}^-$ are drawn independently from the distribution.

TABLE I

FEATURES EXTRACTED FROM VISUAL SENSORY STREAM

Type	Feature	Description
Shape	a, s, b_x, b_y, t	deformable superellipse
Color	R, G, B	RGB color space
Area	A	superellipse area %
Spatial Relation	v_x, v_y	relative orientation

The likelihood function is given by:

$$p(\mathbf{X}^-|\boldsymbol{\theta}^-) = \prod_n \prod_i p(\mathbf{x}^-_{ni}|\boldsymbol{\theta}^-) \tag{1}$$

If we assume a unique Gaussian density, the Maximum Likelihood (ML) solution [2] is:

$$\boldsymbol{\mu}^- = \frac{1}{N} \sum_n \sum_i (\mathbf{x}^-_{ni}) \tag{2}$$

$$\boldsymbol{\Sigma}^- = \frac{1}{N} \sum_n \sum_i (\mathbf{x}^-_{ni} - \boldsymbol{\mu}^-)(\mathbf{x}^-_{ni} - \boldsymbol{\mu}^-) \tag{3}$$

A more difficult problem is to estimate a parametric model $p(\mathbf{x}|c, w, \boldsymbol{\theta}^+)$ from positive bags. As we know, a bag is labeled positive if at least one instance in that bag is positive. However, we do not know which instance is the positive one. The knowledge of positive instance in each bag is modeled by using a set of hidden variables, which are estimated using the Expectation Maximization algorithm. We denote positive bags simply as $\mathbf{x}$, and the ith instance in that bag as $\mathbf{x}_i$. We suppose that each bag have same number of instances, I. We introduce a I-dimensional binary random variable $\mathbf{z}$ having a 1-of-I representation. There are I possible states for the vector $\mathbf{z}$. The value of z_i therefore satisfy $\sum_i z_i = 1$. The hidden variable z models the missing information: the learning process so try to estimate the semantic category and at the same time, approximately what is the proper instance for each bag. We can thus define the likelihood function:

$$p(\mathbf{x}|\mathbf{z}) = \prod_{j=1}^{I} p(\mathbf{x}|z_j)^{z_j} = \prod_{j=1}^{I} p(\mathbf{x}_1, ..., \mathbf{x}_I|z_j)^{z_j} \tag{4}$$

The likelihood function depends only on one of the instances in the bag. We can then rewrite the previous equation as follows:

$$p(\mathbf{x}_1, ..., \mathbf{x}_I|z_j) = \prod_{i=1}^{I} p(\mathbf{x}_i|z_j) \tag{5}$$

At this point, we must quantify the degree of "positivity" of the instance, which depends on two main factors: the instance should not belong to the negative model previously estimated and at the same time it must be like to at least one instance of any other positive example. We can rewrite the equation 5 as follows:

$$p(\mathbf{x}|z_j) = (1 - \aleph(\mathbf{x}_j|\boldsymbol{\theta}^-)) \prod_{i \neq j} \aleph(\mathbf{x}_i|\boldsymbol{\theta}^-) \tag{6}$$

Like Maron's *Diverse Density*, the equation 6 represents a measure of the intersection of the positive bags minus the union of the negative bags [10]. By maximizing that measure, we can find the redundant points distribution (the desired concept). The equation 6 quantifies the possibility that a specific instance of the bag is positive and at the same time depends on the degree of negativity of the other instances, seeking instances farthest from the negative examples, but closer to other positive instances.

$$p(\mathbf{z}) = \phi(\mathbf{z}) = \prod_j \phi(z_j)^{z_j} \tag{7}$$

The ϕ function compute a prior estimate of "positivity" of each instance of a bag. In this work, we do not model this probability (attention). For each bag $\mathbf{x}$ of the training set, we know a I-dimensional real-value vector, $\boldsymbol{\phi}$. The value of ϕ_i satisfy $\sum_i \phi_i = 1$. The posterior probability integrates the salience of the individual instance, which depends on the attention process, with its degree of positivity, which depends on the entire training set and is therefore more generic. The posterior probability is defined as follows:

$$p(z_j|\mathbf{x}) = \frac{\phi_j p(\mathbf{x}|z_j)}{\sum_{i=1}^{I} \phi_i p(\mathbf{x}|z_i)} \approx \phi_j p(\mathbf{x}|z_j) \tag{8}$$

Now consider the problem of maximizing the likelihood for the complete data set $\mathbf{X}, \mathbf{Z}$:

$$\ln p(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta}^+) = \sum_n \sum_i z_{ni} \ln \aleph(\mathbf{x}_{ni}|\boldsymbol{\theta}^+) \tag{9}$$

During expectation step, we estimate the expected value of the variable z_{ni}.

$$p(\mathbf{Z}|\mathbf{X}, \boldsymbol{\theta}^+) \approx \prod_n \prod_j \left[\phi_{nj} p(\mathbf{x}_n|z_{nj})\right]^{z_{nj}} \tag{10}$$

$$E\left[z_{ni}\right] = \frac{\phi_{ni} p(\mathbf{x}_n|z_{ni})}{\sum_{j=1}^{I} \phi_j p(\mathbf{x}_n|z_{nj})} = \gamma(z_{ni}) \tag{11}$$

We can now proceed as follows.

$$\boldsymbol{\mu}^+ = \frac{1}{N} \sum_n \left[\max_i \gamma(z_{ni})\right] \mathbf{x}_{ni} \tag{12}$$

$$\boldsymbol{\Sigma}^+ = \frac{1}{N} \sum_n \left[\max_i \gamma(z_{ni})\right] (\mathbf{x}_{ni} - \boldsymbol{\mu}^+)(\mathbf{x}_{ni} - \boldsymbol{\mu}^+) \tag{13}$$

$$N = \sum_n \left[\max_i \gamma(z_{ni})\right] \tag{14}$$

Instead of using all the instances of each positive bag for density estimation, we use only the instance that maximizes the expected value on the hidden variable. In this way, we consider the hypothesis made initially, namely that each positive bag contains only one positive instance.

B. Word-to-meaning association

In the previous section we presented the algorithm used for the estimation of parametric distributions that describe the positive and negative instances for each semantic class. We must now associate each word with an estimated semantic category, and then force the system to make the more correct association. We can apply our previous algorithm and evaluate the degree of association between semantic class and

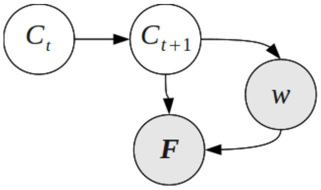


Fig. 2. Word-to-meaning association: probabilistic model

word. We use a distortion measure (Bhattacharyya distance) between the positive and negative distributions as a measure of association between word and semantic categories. Once we have determined the semantic category that maximizes the probability measure for each word (treated as the more correct association), we can estimate a pseudo syntax that generalizes the results obtained. Exploiting only the sensory information to determine the word-to-meaning association still leads to partially correct results. Densities used are in fact estimated from extremely noisy data, and the association procedure will be successful only for those words whose training set is less ambiguous. We must try to integrate into the process other information that will enable us to correct those ambiguous cases.

One possibility is to add more examples in those training sets which are corrupted by more noise, and re-apply the learning algorithm. Another option is to use the position of words in the sentence along with the semantic information to improve the associations: words associated with similar perceptual categories will be included in same syntactic category and will follow same syntactic rules. But we must first solve the problem of the referent, which remains unknown. Our previous learning algorithm allows us to approximate partially correct language model that can be used together with the salience to determine the object of interest in the description. In some cases this process might fail by selecting the wrong referent. We have defined a fitting function which measures the similarity of an utterance to an object based on Mahalanobis distance and saliency. The acquired lexical items allow the parser to assign a semantic category (and hence a Gaussian density) to each word of the sentence. The fitting function is defined as:

$$\frac{\sum_{i=1}^{T}\sqrt{(f-\mu_i)\Sigma_i^{-1}(f-\mu_i)}}{\phi} \tag{15}$$

The object of the scene that minimizes this measure is selected as a possible referent of the sentence. This process can be repeated each time the language model change. This process is depicted in Fig. 2. Each example of the (positive) training set then consists of a sequence of words $w_{1:T_k}$ (utterance) and a set of features $F_k = f_k^1, \ldots, f_k^M$ describing the alleged target object (M feature type or semantic class). C latent variable models the relationship between word w and one of the feature classes f^m, and is time-dependent. We want to find the sequence of semantic classes (hidden variable) that gave rise to a certain sequence of words and a given set of features (observations). Figure 2 shows the graphical model that summarizes the dependencies between random variables used. The probability $p(C_{1:T_k}|w_{1:T_k}, F_k)$ can be decomposed in a manner similar to HMMs, as follows:

$$p(C_0)\prod_{t=1}^{T_k} p(C_t|C_{t-1})p(w_t|C_t)p(F_k|C_t,w_t) \tag{16}$$

As in Hidden Markov Models (HMM), we recognize the transition probabilities of semantic categories $\mathbf{A}$, the emission probability of the word given a particular semantic category $\mathbf{B}$, and the probability of the features given the semantic category and word. This last probability coincides with the semantic category $p(\mathbf{x}|c, w, \boldsymbol{\theta}^+)$ estimated at the previous step for each class-word pair.

$$p(F_k|C_t = m, w_t = i) = \aleph(f_k^m|\boldsymbol{\mu}_{mi}^+, \boldsymbol{\Sigma}_{mi}^+) \tag{17}$$

The discrete variable C can take M values, while w can take W values. We want to estimate the parameter of the model $\theta = \{\mathbf{A}^{M\times M}, \mathbf{B}^{M\times W}\}$. The transition probability $\mathbf{A}$ encodes a pseudo-syntax that depends on the semantic classes. The emission probability $\mathbf{B}$ measures the degree of belonging of a word to a particular semantic class. We used a modified version of Baum-Welch algorithm to learn the parameters of the model. In the expectation step, we calculate first the transition from class j to class i given a word s of the sentence and the set of features describing the target object F_k, as follows:

$$\epsilon_t(i,j) = p(C_t = i, C_{t+1} = j|w_t = s, F_k) = \tag{18}$$

$$= \frac{\alpha_t(i)a_{ij}b_{js}\aleph(f_k^j|\boldsymbol{\mu}_{js}^+, \boldsymbol{\Sigma}_{js}^+)\beta_t(j)}{\sum_i\sum_j \alpha_t(i)a_{ij}b_{js}\aleph(f_k^j|\boldsymbol{\mu}_{js}^+, \boldsymbol{\Sigma}_{js}^+)\beta_t(j)} \tag{19}$$

then the probability of being in state i, given the observations sequence and the model:

$$\lambda_t(i) = p(C_t = i|w_t = s, F_k) = \sum_j \epsilon_t(i,j) \tag{20}$$

Maximization with respect to $\mathbf{A}$ and $\mathbf{B}$ is easily achieved by using appropriate Lagrange multipliers with the following result:

$$a_{ij} = \frac{\sum_t \epsilon_t(i,j)}{\sum_t \lambda_t(i)} \tag{21}$$

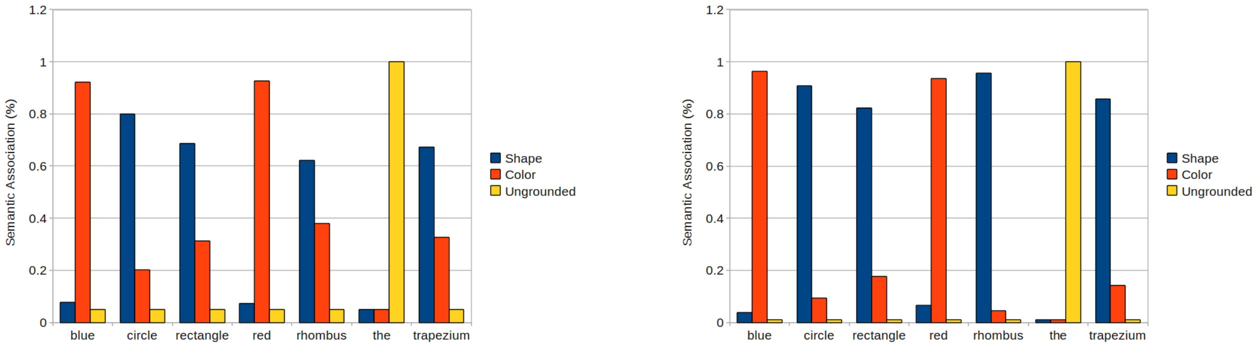
$$b_{js} = \frac{\sum_{t,w_t=s}\lambda_t(j)}{\sum_t \lambda_t(j)} \tag{22}$$

The EM algorithm requires initial values for the parameters of the emission distribution. We can initialize these probability as follows:

$$a_{ij}^0 = \frac{1}{M} \tag{23}$$

$$b_{js}^0 = \frac{bhatt(\boldsymbol{\mu}_{js}^+, \boldsymbol{\Sigma}_{js}^+, \boldsymbol{\mu}_{js}^-, \boldsymbol{\Sigma}_{js}^-)}{\sum_s bhatt(\boldsymbol{\mu}_{js}^+, \boldsymbol{\Sigma}_{js}^+, \boldsymbol{\mu}_{js}^-, \boldsymbol{\Sigma}_{js}^-)} \tag{24}$$

Note that the algorithm is not guaranteed to converge at the global maximum.



(a) Semantic association measured from Gaussian density estimated with multi-instance algorithm

(b) Semantic association obtained after having integrated the syntactic information.

Fig. 3. Results of the word-to-meaning association algorithm. Words associated with similar perceptual categories will be included in same syntactic category and will follow same syntactic rules.

IV. EXPERIMENTAL RESULTS

As previously mentioned, the system has been tested on the NAO robotic platform. NAO is a humanoid robot equipped with Force Sensitive Resistors (FSR) located on the feet, sonars, bumpers, tactile sensors, an IR emitter/receiver, a stereo camera and a pair of microphones. The robot has a number of built-in machine vision modules used in the experimental setup. In addition, we have implemented a set of perceptual and motor schema for basic behaviors such as *pointing* and *grasping*. A typical scene is provided in the Fig. 4.

The experimental setting consists of a set of objects of different shape and color placed on a table. A camera is placed above the table and it ensures a comprehensive view of the scene. Another camera is fixed on the face of the demonstrator and is used for monitoring the gaze direction. The variation of objects is limited to shape, color, size and position. A training corpus from two participants unfamiliar with the project has been collected. The acquisition process is, as already mentioned, interactive: the demonstrator stimulates the attention of the robot on one or more objects of the scene and verbally describes them. Participant were asked to generate simple utterances related to the observed scene such that a listener could later select the same target from the identical scene. Simple utterances contain reference to exactly one object (target object). The training corpus was composed of 266 utterances of which 236 are simple and 30 complex. Only simple utterance are used in learning process. The results of the algorithm are promising. Semantic clustering algorithm is able to isolate more than 80% of positive instances and then to estimate correctly the semantic category associated with the word. Figure 3(left) shows the degree of associability the word calculated in the first phase of the algorithm with respect to semantic classes. In most cases we can still get partially corrected results with our previous algorithm. However there are ambiguities, as in the case of the word ”circle”, which can be minimized by considering syntactic information. Figure 3(right) shows the final results obtained in the second phase of the algorithm.

All concepts underlying acquired language model are used to initialize dynamic fluents as predicate calculus terms and update robot’s knowledge base representing the state of the world from sensor data. Only the actions of the robot can modify the values of the fluent associated with the objects. The demostrator can’t modify the scene. For this reason, the knowledge base is updated after every action of the robot. The only fluent to be updated are those associated with spatial relationships between objects that change with every action. Every time the robot completes the move or grasp actions, updates the database with new spatial relationships. Obviously there will be some of the logic terms that will not vary at all (eg, color). We have tested the capabilities of the robot to understand the descriptions provided by the users and to conduct a dialog in case of ambiguities. The robot was given concrete instructions, such as “*Point the green object!*”, or “*Grasp the object to the left of the yellow circle!*”[1]. The whole human-robot interaction is driven by gestures and language. An example of dialogue is shown below, while the robots actions are depicted in Fig. 4:

Robot: looks at the object 1.
Human: ”NAO, grasp the object to the left of the blue one!”
Human: points the object 1 (50%), 2 (20%), 3 (50%).
Robot: looks at the object 3.
Robot: ”Is it the yellow rectangle?”
Robot: points the object 3 (Fig. 4 left).
Human: ”No!”
Robot: ”Is it the blue circle?”
Robot: points the object 2 (Fig. 4 center).
Human: ”Yes! That’s right!”
Robot: Grasps the blue circle (Fig. 4 right).

We used the disambiguation trees presented in our previous work, to solve some perceptual ambiguities present

[1] In the present model, the meaning of verbs “*to point*” and “*to grasp*” is hand-coded, and it is not learned by the system. Future releases will address the problem of grounding dynamic terms through the same computational framework.

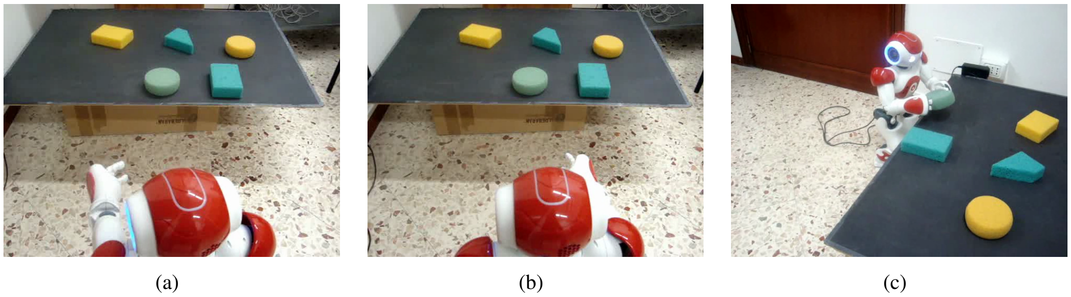


Fig. 4. An example of human-robot interaction via learned language model. The demonstrator asked the NAO to take the object to the left of a blue object. (a) NAO points the first object and says ”Is the yellow rectangle?”; (b) NAO points the second object and says ”Is the blue circle”; (c)NAO grasps the correct object.

in the scene [5]. In this experiment have been learned only the terms that refer to colors and shapes of objects in the scene. For the spatial relationships were used categories learned in our previous work [5]. A set of external observers were judging the goodness of the system with respect to the following factors:

- Naturalness of the robot’s linguistic and motor behavior;
- Differences between the expected behavior and that observed.

About ten people were involved in a full-day evaluation session. The overall score was positive in about 80% of collected forms. While these results have no scientific foundation, they however show a positive impact of our computational model in a human-robot interaction system.

V. CONCLUSION

The algorithms presented in this article extends our previous work on the grounded language model learning. We focused on some limitations of the previous technique while maintaining the same algorithmic structure. In particular: (a) we endowed the system with a real model of attention and formalized a multi-instance learning algorithm for the acquisition of semantic categories, (b) we have improved the word-to-meaning association algorithm, by linking the choice not only to semantic information but also to syntactic constraints encountered, and (c) we have made demonstrator-robot interaction more natural.

However, a set of important questions still remain to be solved. As presented, the system learns “simple” concepts involving a single perceptual channel. Ongoing work is focused on learning complex concepts from the interaction data. The same computational framework will be employed recursively in order to assign meanings to words by hierarchically describing complex concepts as composed of simpler ones in a Bayesian network. Another issue is related to the process of learning and understanding verbs as words that usually involve an observable action. The work presented here represents the first steps in this direction.

VI. ACKNOWLEDGMENTS

This work has been partially supported by the EU funded project HUMANOBS: Humanoids That Learn Socio-Communicative Skills Through Observation, contract no. FP7-STREP-231453 (www.humanobs.org).

REFERENCES

[1] D.A. Baldwin. Understanding the link between joint attention and language. *Joint attention: Its origins and role in development*, pages 131–158, 1995.

[2] C.M. Bishop. *Pattern recognition and machine learning*. Springer New York, 2006.

[3] P. Bloom. Intentionality and word learning. *Trends in cognitive sciences*, 1(1):9–12, 1997.

[4] A. Cangelosi. Approaches to grounding symbols in perceptual and sensorimotor categories. *Handbook of categorization in cognitive science*, pages 719–737, 2005.

[5] H. Dindo and D. Zambuto. Resolving ambiguities in a grounded human-robot interaction. *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium*, pages 408–414, 2009.

[6] C. Fisher, D.G. Hall, S. Rakowitz, and L. Gleitman. When it is better to receive than to give: Syntactic and conceptual constraints on vocabulary growth. *The acquisition of the lexicon*, pages 333–375, 1994.

[7] K. Gold, M. Doniec, C. Crick, and B. Scassellati. Robotic vocabulary building using extension inference and implicit contrast. *Artificial Intelligence*, 173(1):145–166, 2009.

[8] M. Heckmann, H. Brandl, J. Schmuedderich, X. Domont, B. Bolder, I. Mikhailova, H. Janssen, M. Gienger, A. Bendig, T. Rodemann, et al. Teaching a humanoid robot: Headset-free speech interaction for audio-visual association learning. *The 18th IEEE International Symposium on Robot and Human Interactive Communication, 2009. RO-MAN 2009*, pages 422–427, 2009.

[9] G. Herzog and P. Wazinski. VIsual TRAnslator: Linking perceptions and natural language descriptions. *Artificial Intelligence Review*, 8(2):175–187, 1994.

[10] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems*, pages 570–576. MIT Press Cambridge, MA, USA, 1998.

[11] D. Roy. Learning visually grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3):353–385, 2002.

[12] D. Roy. Grounding Words in Perception and Action: Insights from Computational Models. *Trends in Cognitive Science*, 9(8):389–396, 2005.

[13] L. Steels and F. Kaplan. Bootstrapping grounded word semantics. *Linguistic evolution through language acquisition*, page 53, 2002.