Visually-Grounded Language Model for Human-Robot Interaction

Daniele Zambuto and **Haris Dindo** and **Antonio Chella** [1]

Abstract. Visually grounded human-robot interaction is recognized to be an essential ingredient of socially intelligent robots, and the integration of vision and language increasingly attracts attention of researchers in diverse fields. However, most systems lack the capability to adapt and expand themselves beyond the preprogrammed set of communicative behaviors. Their linguistic capabilities are still far from being satisfactory which make them unsuitable for real-world applications. In this paper we will present a system in which a robotic agent can learn a grounded language model by actively interacting with a human user. The model is grounded in the sense that meaning of the words is linked to a concrete sensorimotor experience of the agent, and linguistic rules are automatically extracted from the interaction data. The system has been tested on the NAO humanoid robot and it has been used to understand and generate appropriate natural language descriptions of real objects. The system is also capable of conducting a verbal interaction with a human partner in potentially ambiguous situations.

1 Introduction

The aim of this work is to investigate the lexical acquisition problem, namely how can a robot be bootstrapped into communication and what are the necessary prerequisites for robots in order to learn a language? In particular, the focus is set on grounded systems that learn to generate and understand contextualized spoken descriptions of objects in visual scenes. There are two basic problems the robots need to solve as they acquire a language:

1. Identify the meaning of words grounded in perceptual data;
2. Infer a rudimentary grammar for further understanding and interaction.

The process of lexical acquisition in infants seems to be innately driven by the principle of reference: words refer to objects, actions, and attributes of the environment. Observational learning may be used to deduce word meanings from cross-situational experiences. A well-known problem in observational learning, is the Quine's paradox: an infinite number of possible meanings can be inferred from a finite set of utterance-context pairs. A likely solution to this problem is that all infants have certain biases which constrain the set of possible meanings of words [3, 1]. For example, the whole object assumption proposes that children will assume a novel label refers to a whole object rather than its parts. The mutual exclusion assumption proposes that they prefer to assign only one label to a concept. These assumptions are considered good strategies for bootstrapping the inference process.

Present approach uses these assumptions to bootstrap the lexical acquisition process. 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. Initially, the system will acquire a minimal grounded language model from the paired *description-referent* examples, without any prior semantic and syntactic information, except the main referent of description utterance, from which the system will be able to acquire novel concepts and more complex language models (e.g., spatial clauses).

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. 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. Language acquisition therefore proceed in parallel with concept acquisition. Concepts acquired from lexical acquisition are used to initialize a logic representation of several observable entity of world. For example, the meaning of the word "red" is used to seed the logic representation of the "red" concept. Concepts underlying acquired language model can be considered as independent from language acquisition process and can be reused for other cognitive tasks. Fig. 1 provides an overview of the learning system.

The system has been tested on the NAO humanoid robot and it has been used to understand and generate appropriate natural language descriptions of real objects. The system is also capable of conducting a verbal interaction with a human partner in potentially ambiguous situations, e.g. when a human interlocutor refers to several co-occurring entities (i.e. objects sharing identical perceptual properties) in the world.

2 Related works

There has been a huge interest in grounded language acquisition in the past years. Probably, the earliest and the most influent work on language analysis is Winograd's SHRDLU [11] in which an artificial agent connected to a simulated robotic arm is capable of understanding questions on the world, execute given actions, and ask for help in case of ambiguous dialogue. However, SHRDLU is based on a purely symbolic representation of the world, and the artificial agent has no cognition about the "real" meaning of a given sentence. It is mostly based on a pure syntactic analyzer, and the semantics is deeply hardwired into the system. Visual Translator system [6] (VITRA) is a natural language generation system which is grounded

[1] University of Palermo, Computer Science Engineering, Viale delle Scienze, Ed. 6, Palermo, Italy, email: {dindo, zambuto}@dinfo.unipa.it, chella@unipa.it

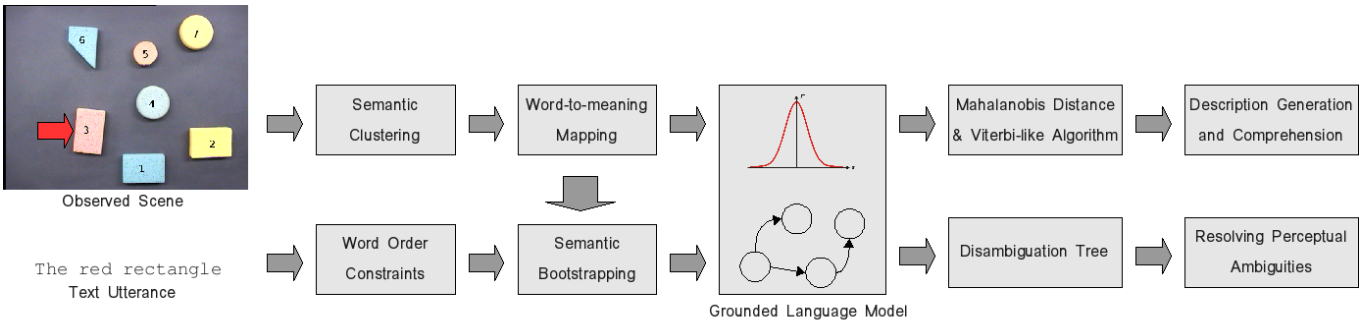


Figure 1. Model of grounded language acquisition: from word learning to simple language and dialogue

directly in perceptual input. From a sequence of digitized video frames low-level sensory processes perform recognition and tracking of visible objects. Authors provide a geometrical reconstruction of the perceived scene. 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. Roy implemented a system, CELL [9] (Cross-channel Early Lexical Learning), able to learn object names from a corpus of spontaneous infant-directed speech, and to process single and two-word phrases which referred to the colour 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. In DESCRIBER [8], the same authors address 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 [4], 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 word 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. In another related approach, authors develop language acquisition system without knowledge of grammar and vocabulary, which learns concrete noun concepts from user utterances and paired images [10]. The system adopts a frequentist approach, in which utterances and images containing the same objects are processed and the most probable noun is chosen as label.

Our work, while not making significant advances compared to the systems at the state-of-the-art, puts more emphasis on a fundamental problem in the language acquisition process, namely the selection of the referent. Our goal is to create robust grounded language learning algorithms, able to build operational knowledge even in absence of important pragmatic information.

3 Model overview

Before going into details of 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 colour spectrum. Such a semantic category could be used to ground the semantics for a colour 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 colour terms (colour class). Finally, a *lexical item* encodes the association between a word and its corresponding semantic category.

In the experimental setting each training sample is comprised of an utterance and a visual context representing the semantics of the word sequence. Utterances consist of phonetic transcripts of spoken sequences recorded by the auditory sensor. Context consists of visual sensory input which co-occurs with the utterance, and usually it contains instances of multiple semantic class. A key assumption is that any word in the utterance may refer to any semantic category inferred from the co-occurring context.
All semantic categories are derived from visual sensory signals. 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, colour, size, and spatial relation. The context is initially restricted to a given referent (target) object, and any word may potentially be paired with any semantic category which is derived from the same utterance-context pair. A typical teaching scenario is depicted in Fig. 1 (left).

3.1 Learning how to ground words to perceptual stimuli

Grounding can be considered as the process whereby internal representations are connected to external percepts. This is based on the principle that cognitive agents and robots learn to name entities, individuals and states in the external (and internal) world at the same time as they interact with their environment and build sensorimotor representations of it [5]. Our approach addresses two interrelated question in lexical acquisition: how semantic categories can be learned and

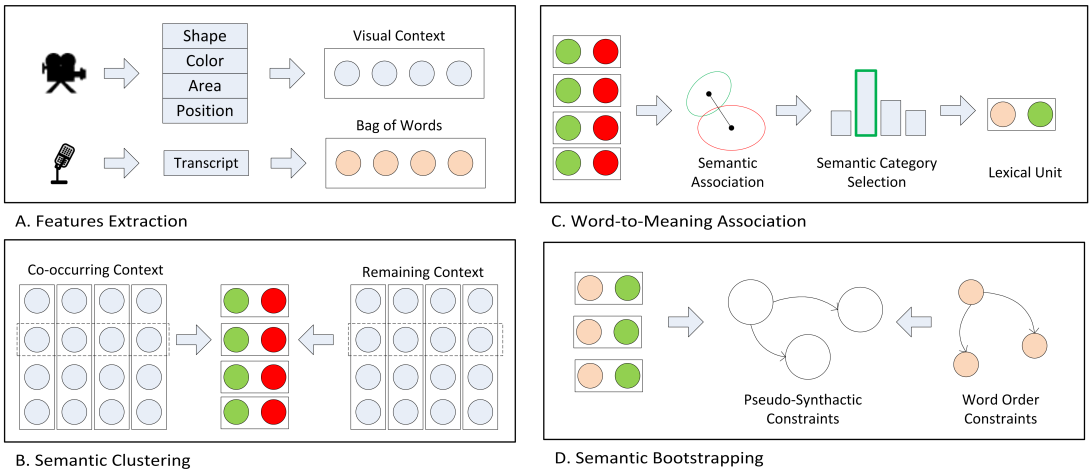


Figure 2. Computational processes composing our model for language acquisition

how symbols/words can be associated with appropriate semantic categories. These problems are known as semantic categorization (or *semantic clustering*) and *word-to-meaning mappings* respectively. No *a priori* knowledge about innate semantic categories is assumed by the model; instead, through repeated experience, appropriate categories must be learned from positive examples and the model must associate linguistic units with appropriate semantic categories. This process is depicted in Fig. 2.

Basic problem regarding semantic clustering is how to establish the semantics of individual words since each word in an utterance may potentially be associated with any subset of co-occurring visual features. For the sake of simplicity, we can assume that each word may be associated with only one of co-occurring semantic categories. However, natural language usually does not provide exhaustive labels of all referents in a scene. In other words, we cannot assume that the absence of a word indicates the absence of the corresponding property. As a consequence, appropriate categories must be learned from the positive examples only.

Semantic clustering generates the possible meanings of words captured by the system. This first step is not concerned with associating a word to a particular meaning, but it deals with the problem of finding *all* the possible meanings given the training set. The meaning of each word (i.e. its semantic category) is treated as a random variable and modeled with a multivariate Gaussian distributions. For example, the meaning of a word associated to the semantic class 'color' can be associated with a multivariate Gaussian distribution over a red-green-blue colour space. Each semantic category is composed of two parametric distribution: the distribution of feature values conditioned on the presence of word $p(f^j|w)$ (hypothetical semantic category) and the distribution of feature values conditioned on the absence of word $p(f^j|\overline{w})$ (background distribution). Only the visual features f which occurred with the word w are used to compute the unbiased estimates of the word conditioned model. The remaining observations (i.e. visual context of examples in which the word is not present) are used to estimate the background model. These distributions are estimated for each semantic class C_j (shape, colour, size, spatial relation). This process is represented in Fig. 2(B).

These densities represent a possible meaning of the word. A semantic distortion metric is used to select appropriate semantic category from several hypothetical ones. We use the distortion between the word-conditioned and background distributions as a measure of association between word and semantic categories. The Bhattacharyya distance measures the similarity of two probability distributions, and the optimal Bayes classification error between any two classes can be bounded by the Bhattacharyya error, ε. In our case, the Bhattacharyya error provides a measure of association between words and individual semantic categories. A clustering algorithm estimates hypothetical semantic categories of words from co-occurring contexts and associates each word with semantic categories that maximize the classification error outlined above [2]. Categories below a fixed threshold are discarded, and the word is associated to a special ungrounded class. Ungrounded class gathers all the words that can not be directly grounded in sensory input. In our model, no semantic category (Gaussian density) is associated with these words. The whole process, together with the definition of the involved quantities, is depicted in the Fig. 2(B-C).

3.2 Grammar learning

Categories, such as adjective, verbs and nouns, form the basic units for learning the rules of grammar including grammatical relations, cases, and phrase structure configurations. Without syntactic categories a learner will be unable to acquire the rules of the language. Bootstrapping theories provide strategies for deriving syntactic categories from perceptual data. For instance, semantic bootstrapping [7] proposes that the language learner uses semantic categories to seed syntactic categories. This theory assumes that the learner has already acquired words and their semantics without the use of any syntax.

Basic syntactic categories will be acquired directly through words' meaning, and the words associated with similar perceptual categories will be included in same syntactic category. The grammar is modeled as a Markovian chain. We collect word statistics from utterance training corpus encoding: (1) the word transition probability, (2) the probability of beginning an utterance with a word and (3) the probability of ending an utterance with a word. Lexical items generated in the previous phase can be used to cluster words into groups that depend on the associated meaning. We use these groups, and associated

[2] We force the clustering algorithm to associate only one semantic category, without considering multi-class meanings (semantic categories that depend on several classes of perception)

probabilities, to construct a more general representation of syntactic constraints in terms of a finite state automaton (FSA) based on semantic classes. These FSA are used as the basis for a deterministic parser which identifies object phrases embedded in complex utterances.

3.3 Learning complex spatial concepts

However, complex adjectives, such as "above" and "below", cannot be learned in the same fashion. Instead, we need to take the order of words into account. Treating sentences as unordered set of words tends to lose syntactic contents and its semantic implications. We need to exploit the word ordering constraints and semantic bootstrapping assumption to acquire spatial terms and explore weak generalization mechanisms.

To learn the meaning of spatial terms, we need to find all the objects to which the phrase refers. We consider only those phrases describing spatial relationships between two objects (i.e., "the rectangle above the red circle"). A deterministic parser based on previously acquired FSA, identifies object phrases embedded in complex utterances. We must decide which of two phrases refers to the original target object, and which of the remaining objects in the scene should be linked to the other phrase. We have defined a fitting function which measures the similarity of an utterance to an object based on Mahalanobis distance, which will be explained later. Once the focal objects of the sentence has been determined, we can calculate the features that describe the spatial relationship between these. We must also encode the order of the phrases to learn the difference between spatial terms (i.e., difference between "above" and "below") and to calculate correctly the spatial features. The parser replace the object phrases with (ungrounded) labels to facilitate the task. The only information that really matters in our case, is which of the two sentences refers to the target object. Finally, the learning process that associate a meaning to the word describing a spatial feature continues as previously described. New lexical items are simply added to the vocabulary of model, and the learned FSAs are merged into a single syntactic model.

3.4 Understanding and generating descriptions

Having learned the probabilistic grounded language model, the task of understanding an utterance can be casted in terms of classical operations on probabilistic graphical models. Thanks to the Mahalanobis distance, we can measure the degree of association ("semantic distance") between a visual feature and a semantic category previously acquired. For example, we consider the case of a sentence like "the red rectangle". The acquired lexical items allow the parser to assign a semantic category (and hence a Gaussian density) to each word of the sentence. Given the visual features of an object f, we can calculate the degree of association between sentence (of length T) and object's visual features as:

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

The object of the scene that minimizes this measure is selected as a possible referent of the sentence. The same procedure can be used for understanding spatial clauses. The deterministic parser guides the extraction of spatial features which depends on the order of object phrases. The referent of the sentence is searched among all possible pairs of objects.

One way to facilitate the understanding of descriptions and to decrease the computational burden is to use logical terms to represent the state of the world (i.e., knowledge about the properties and spatial relationships of objects), and then use query in a logic programming language for inferring the reference object. When the parser begins to analyze an utterance, it queries the system for the state of the world described in the form of a list of all atomic sentences that can be produced from available sensors at the time of query. For each object in the scene, some logical terms are generated through the acquired lexical items. In particular, the system generates a logical term for each visual feature extracted from a object by selecting the lexical item that minimizes the Mahalanobis distance for each semantic class. Deterministic parser has been modified to generate logical query to be submitted to the system. Some examples of possible queries are: "red(X)", "rect(X), above(X,Y), red(Y)", etc.

The generation of the description of an object is quite simple. We can generate a description as the most likely path of words (in the main FSA) that produces a given observation. Given a set of visual features extracted from an object we generate a description of object through a modified Viterbi-like algorithm. We seek the sequence of words that minimizes the fitting function for that object, and follows the syntactic rules implicitly encoded in the FSA. We need a method to determine the length of a description. The generation algorithm include a stop criterion based on the acquired ending probability.

4 Dialogue and ambiguities

Another problem we have addressed is how to resolve the ambiguities contained in a description through simple verbal interactions. For example, suppose that the robot hears the statement "Grasp the small blue object!" while facing a scene containing several objects having the same perceptual features.

In order to fulfill the task, the robot must disambiguate the context and select one of the referred objects. The process of ambiguity resolution is based on a context-specific human-robot dialogue through specific questions directly related to the properties and spatial relationships of the objects in the observed scene. We have considered only those ambiguities related to perceptual quality shared between objects of the scene. The robot must be able to detect ambiguous statements and select the less ambiguous question to be asked in order to succesfully conduct a dialogue. For instance, if all red objects have different shapes, the robot can easily resolve the ambiguities by asking for the shape of a specific object, i.e. "Are you referring to a rectangle?". However, if all items have the same shape, the robot must solve the ambiguity by using relative spatial terms between objects that are usually less informative than the concrete shape-based questions. For example, many objects may share particular spatial relationships with other objects making the question partially ambiguous. In addition, the landmark object may be indistinguishable from the target one making the phrase even more ambiguous.

Possible ambiguities can be removed through a simple dialogue based on yes/no questions. *Disambiguation tree* reconstruct the robot's decision process in choosing a question to disambiguate a context. They are essentially decision trees: the possible target objects are stored in the leaves, a disambiguation strategy is given by the path from the root to the object's leaf. The interior nodes can be questions about referent's proprieties itself, or relations to other objects. When the robot finds a reference ambiguity, it builds a disambiguation tree that resolves it. Fig. 3 shows an example of disambiguation tree. Path to the left after question indicate that the predicate is satisfied (yes answer), while the right branch indicates that

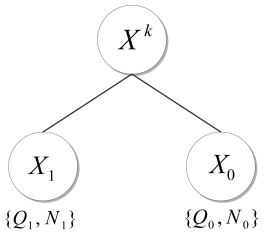


Figure 3. Disambiguation tree represents the underlying decision process of choosing a question able to disambiguate a visual context

it is not (no answer). Each question consists of attempting to satisfy a logical predicate, and appropriate subtree is explored. This process continues until a leaf is reached and alle the ambiguities are resolved.

Disambiguation trees are constructed using the output of an ambiguous query (logic queries with more than one possible outcome). The construction method minimizes an entropy-based disambiguation measure. At each decision node, the algorithm chooses a possible question and splits the available objects into two groups. This process then occurs recursively until all objects are selected. Clearly, the algorithm must decide which of these question is most informative. We have used a entropy-based disambiguation measure, characterizing the average amount of uncertainty in a single question. This measure is computed as:

$$\frac{N_0}{N_1}H(X_0) + \frac{N_1}{N_0}H(X_1) + \sum_i C_i(X^k)$$

The first two members of the equation measures the uncertainty related to possible questions in the children nodes (future scenarios). We have calculated the uncertainty linked to the set of possible questions Q in a node, as the entropy calculated on a random variable X generated by this set. The algorithm selects one of the possible questions associated with the current node (represented by X^k) and split the set of objects into two groups. For each group generates all the possible questions Q_i ($i = \{0, 1\}$) that have not been previously selected in other high level decision node. The discrete random variable X_i encodes the information related to each question in set Q_i. We can compute a probalility mass function $p(X_i)$ as follows:

$$p(X_i = j) = \frac{< j >}{\sum_k p(X_i = k)}$$

where the counter operator $< ... >$ return the number of the objects that respond positively to that question. N_i represents the number of the objects of each group. We weight the first two terms by their ratio to ensure that the tree is well-balanced compared to the classical decision tree learning algorithms [2].

The last term measures the uncertainty related to the question selected, depending on the encoding of the question in natural language (i.e. the number of words used) and its contextual ambiguity (i.e. ambiguity in the understanding of the sentence by a listener). The greedy algorithm selects the question that minimizes this measure based on entropy. The trees generated by the disambiguation algorithm, are well-balanced. Fig. 4(b) shows an example of disambiguation tree generated by the algorithm. In this example, the first selected question is the term "blue(X)". The question "Is blue?" allows the system to divide the objects of the scene into two groups: blue objects and not blue objects. The construction method consider future scenarios, and hence possible question and ambiguities in future contexts, to select appropriate question. The method does not select the questions that have an immediate and obvious answer (i.e. "trapezium(X)"), but more general questions that allow us to isolate and partially resolve the perceptual ambiguity and reduce the number of questions to ask. The subsequent scenarios, as mentioned, are those with a lesser degree of ambiguity.

5 Experiments

We have performed two set of related experiments in order to test the proposed model. One set involves a quantitative evaluation of the model in the task of describing a scene. Another set is mostly qualitative and involves an experiment on the NAO robotic platform. Both experiments will be described in the following

Table 1. Results of an evaluation of human and machine generated descriptions.

Participants	Generated by human	Generated by system
A	87,3 %	80,5 %
B	85,5 %	79,8 %
C	88,3 %	81,3 %
Media	87,1 %	80,5 %

Table 2. Results of an evaluation of machine understanding capabilities.

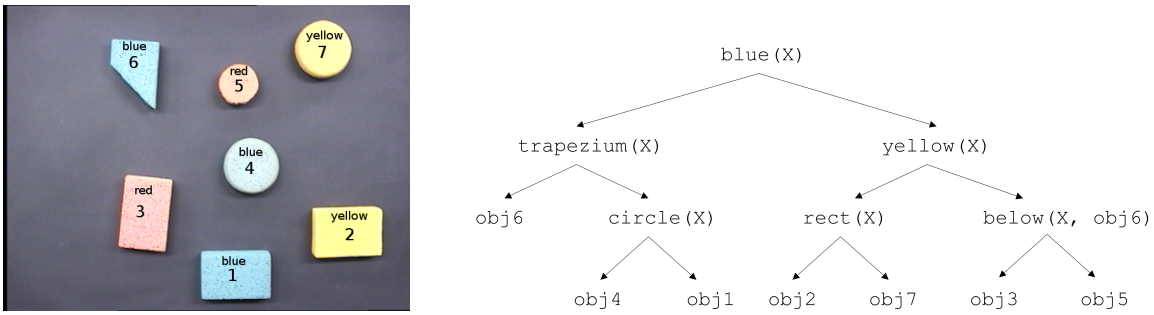
Participants	Training corpus	Testing corpus
A	90,9 %	87,3 %
B	88,7 %	85,5 %
C	91,1 %	88,3 %
System	86,2 %	78,9 %

5.1 Object description: quantitative results

The description task consists of generating phrases which best describe target objects and must be context sensitive since often depend on the other objects in the observed scene. The variation of objects is limited to shape, color, size and position. Each training example is comprised of an utterance and a context representing the semantics of the word sequence. Utterances consist of phonetic transcripts of spoken sequences recorded by the auditory sensor. Context consists of visual sensory input which co-occurs with the utterance, and usually it contains instances of multiple semantic categories. The experimental setting consists of a set of objects of different shape and color placed on a table. The camera is placed above the table and it ensures a comprehensive view of the scene.

A training corpus from two participants unfamiliar with the project has been collected. Participant were asked to generate two different utterances related to the observed scene such that a listener could later select the same target from the identical scene with the target unmarked. Simple utterances contain reference to exactly one object (target object), whereas complex utterances describe two or more objects and their spatial relations. The reference object of a spatial relation was selected by participants during the data collection task. The training corpus was composed of 356 utterances of which 196 are simple and 160 complex.

Participants were asked to select the object which best fit the description generated by the system. All participants evaluated the same sets of images. Responses were evaluated by comparing the selected object for each image (the target object described by the system) to the actual target object which was selected by participant as



(a) A sample image from training corpus

(b) A disambiguation tree inferred from the scene

Figure 4. Disambiguation trees are used to resolve the ambiguities contained in a description through simple yes/no questions; (a) an example of ambiguous context and (b) the corresponding well-balanced disambiguation tree constructed by minimizing an entropy-based measure - each node represents a question to ask in order to conduct a dialogue

the referent of the description. The collected data was used to measure the accuracy of the system-generated description (showed in Table 1). Participants were also asked to select a landmark object and to generate a detailed description of the target object in the scene. Responses were evaluated by comparing the selected objects for each image to the actual target object and landmark object which were selected by the system. Table 2 shows the result of evaluation of system understanding capabilities.

The system was able to describe scenes not encountered during the training phase, and to exhibit sequences of words which have never occurred in the training data. The results presented in this section demonstrate the effectiveness of the learning algorithms to acquire and apply grounded structures for the visual description task.

5.2 Human-robot interaction: qualitative 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`.

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. For this reason, the knowledge base is updated after every action of the robot. Obviously there are several logic terms that are constant over time (e.g. color). We have tested the capabilities of the robot to understand the descriptions provided by the users and to conduct a dialogue 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!*"[3]. The previously described disambiguation trees were used to clarify eventual ambiguities induced by the dialogue. An example of dialogue is shown below, while the robots actions are depicted in Fig. 5:

[3] 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.

Human: "NAO, grasp the object on the left to the blue one!"
Robot: Points the yellow rectangle.
Robot: "Is that the yellow rectangle I am pointing at?"
Human: "No!"
Robot: Points the blue circle.
Robot: "Is that the blue circle I am pointing at?"
Human: "Yes! That's right!"
Robot: Grasps the blue circle.

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 and observed behavior.

About ten people were involved in a full-day evaluation session. The overall score was positive in about 80% of collected forms.

6 Conclusions and future works

In this paper we have presented a computational model for the acquisition of a grounded language model to be used in human-robot interaction. The system learns the meaning of the words and their grammatical usage while understanding or generating a sentence. A related computational model constructs a disambiguation tree in order to help the robot conducting a dialogue in real-world settings.

While the system has been tested using english language, the generality of our computational model makes it an ideal candidate for learning grounded models of other natural languages, provided that every concept is linked with exactly one meaning and that Markov models can be used to represent its (simplified) grammar.

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.

Figure 5. An example of human-robot interaction via the learned language model

ACKNOWLEDGEMENTS

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*, 131–158, (1995).

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

[3] 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*, 333–375, (1994).

[4] 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).

[5] S. Harnad, 'The symbol grounding problem', *Physica. D*, **42**(1-3), 335–346, (1990).

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

[7] S. Pinker, *Language learnability and language development*, Harvard University Press Cambridge, Mass, 1984.

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

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

[10] Y. Uchida and K. Araki, 'Evaluation of a system for noun concepts acquisition from utterances about images (SINCA) using daily conversation data', in *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pp. 65–68. Association for Computational Linguistics, (2009).

[11] T. Winograd, *Understanding natural language*, Academic Press, Inc. Orlando, FL, USA, 1972.