A Discriminative Approach to Grounded Spoken Language Understanding in Interactive Robotics

Emanuele Bastianelli[1], **Danilo Croce**[2], **Andrea Vanzo**[3], **Roberto Basili**[2], **Daniele Nardi**[3]
[1]DICII, [2]DII, University of Rome Tor Vergata, Rome, Italy
[3]DIAG, Sapienza University of Rome, Rome, Italy
bastianelli@ing.uniroma2.it, {croce,basili}@info.uniroma2.it, {vanzo,nardi}@dis.uniroma1.it

Abstract

Spoken Language Understanding in Interactive Robotics provides computational models of human-machine communication based on the vocal input. However, robots operate in specific environments and the correct interpretation of the spoken sentences depends on the physical, cognitive and linguistic aspects triggered by the operational environment. Grounded language processing should exploit both the physical constraints of the context as well as knowledge assumptions of the robot. These include the subjective perception of the environment that explicitly affects linguistic reasoning. In this work, a standard linguistic pipeline for semantic parsing is extended toward a form of perceptually informed natural language processing that combines discriminative learning and distributional semantics. Empirical results achieve up to a 40% of relative error reduction.

1 Introduction

End-to-end communication processes in natural language are challenging for robots for the deep interaction of different cognitive abilities. For a robot to react to a user command like "*take the book on the table*" a number of implicit assumptions should be met. First, at least two entities, a `book` and a `table`, must exist in the environment and the speaker must be aware of such entities. Accordingly, the robot must have access to an inner representation of the objects, e.g. an explicit map of the environment. Second, mappings from lexical references to real world entities must be available. *Grounding* here [Harnad, 1990] links symbols (e.g. words) to the corresponding perceptual information.

Spoken Language Understanding (SLU) in interactive dialogue systems acquires a specific nature, when applied in Interactive Robotics. Linguistic interactions are context aware in the sense that both the user and the robot access and make reference to the environment (i.e. entities of the real world). In the above example, "*taking*" is the intended action whenever a book is actually on the table, so that *the book on the table* refers to the whole argument. On the contrary, the command may refer to a "*bringing*" action, when no book is on

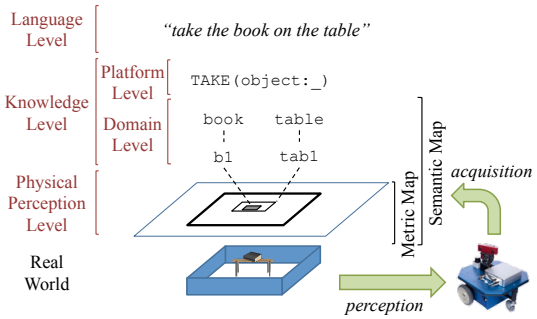

Figure 1: Levels of representation in interactive robotics

the table and *the book* and *on the table* correspond to different arguments. Robot interactions need thus to be *grounded*, as meaning must correspond to the physical world and interpretation is strongly interlaced with what is perceived, as pointed out by psycho-linguistic theories [Tanenhaus *et al.*, 1995]. As a consequence, a correct interpretation is more than a linguistically motivated mapping from an audio signal (e.g. the spoken command) to a meaning representation formalism compatible with a linguistic theory (e.g., semantic frames as discussed in [Fillmore, 1985]). Correctness implies a physical coherence, as entities in the environment must be known and the intended predicates must correspond to (possibly known) actions coherent with the environment too.

While traditional SLU mostly relies on linguistic information contained in texts (i.e., derived only from transcribed words), its application in Interactive Robotics depends on a variety of other factors, including the perception of the environment. We can organize these factors into a layered representation as shown in Figure 1. First, we rely on the *language level* that governs linguistic inferences: it includes observations (e.g. sequences of transcribed words) as well as the linguistic assumptions of the speaker, here modeled through frame-like predicates by which the inner lexicon can be organized. Similarly, evidences involved by the robot's perception of the world must be taken into account. The physical level, i.e. the real world, is embodied in the *physical perception level*: we assume that the robot has an image of this world where the existence and the spatial properties of entities are represented. Such representation is built by mapping the direct input of robot sensors into geomet-

rical representations, e.g. *metric maps*. These provide a structure suitable for anchoring the *knowledge level*. Here *symbols* (i.e., knowledge primitives) are used to refer to real world entities and their properties inside the *domain level*. This comprises active concepts the robot sees realized in a specific environment, plus general knowledge it has about the domain. All this information plays a crucial role during linguistic interactions. The integration of metric information with notions from the knowledge level provides an augmented representation of the environment, called *semantic map* [Nüchter and Hertzberg, 2008]. In this map, the existence of real world objects can be associated to *lexical* information, in the form of entity names given by a knowledge engineer or spoken by a user for a pointed object, as in Human-Augmented Mapping [Diosi *et al.*, 2005; Bastianelli *et al.*, 2013]. It is worth noticing that the robot itself is a special entity described at this knowledge level: it does know its constituent parts as well as its capabilities, that are the actions it is able to perform. In our case, we introduce an additional level (namely *platform level*), whose information is instantiated in a knowledge base called *Platform Model*. In this way, a more complete perceptual knowledge level is accessible, comprising both a model of the world and a model of the robot itself. While SLU for Interactive Robotics have been mostly carried out over the evidences specific to the linguistic level, e.g., in [Chen and Mooney, 2011; Matuszek *et al.*, 2012b; Bastianelli *et al.*, 2014a], we argue that such process should deal with all the aforementioned layers in an harmonized and coherent manner. All linguistic primitives, including predicates and semantic arguments, correspond to perceptual counterparts, such as plans, robot's actions or entities involved in the underlying events.

The main contribution of this work is a SLU process that depends not only on the linguistic information, but also on the perceptual knowledge level. This process is expected to produce interpretations that coherently mediate among the world (with all the entities composing it), the robotic platform (with all its inner representations and its capabilities) and the pure linguistic level triggered by a sentence. To this end, a discriminative approach to SLU has been adopted, where grounded information is directly injected within the learning algorithm, showing that the integration of linguistic and perceptual knowledge improves the quality and robustness of the overall interpretation process. Such goal has been achieved by feeding the learning algorithms with a representation which is enriched with perceptual knowledge extracted from a semantic map, through a grounding mechanism based on linguistic evidences.

In the remaining of the paper, after a short survey on related works (Section 2), we will describe and empirically evaluate the proposed approach in Section 3 and 4, respectively.

2 Related Works

The interest in Spoken Language Understanding (SLU) for Interactive Robotics has grown in the last decade. Researchers with different backgrounds have proposed approaches to the problem, adopting techniques either drawing from traditional Natural Language Processing, or specifically designing both algorithms and representation formalisms.

In [Bos and Oka, 2007], SLU is performed in an integrated fashion with the Automatic Speech Recognition (ASR), by augmenting recognition grammar rules with semantic attachments. Final interpretations are produced together with the utterance transcription, and given according to a logic formalism based on λ-calculus. Final grounding is performed through a theorem prover. Contextual Categorial Grammars (CCG) are used in [Kruijff *et al.*, 2007] to parse transcriptions obtained through ASR. Statistical Learning techniques have been applied to train specific semantic parsers. In [Kollar *et al.*, 2010], Conditional Random Fields are used to sequentially label sentences to extract Spatial Description Clauses (SDCs), i.e. structures describing spatial relations. A maximum-likelihood approach is used to infer the path to follow according to the spatial constraint induced by the SDCs. In [Chen and Mooney, 2011], language descriptions are paired with robotic actions, and Statistical Machine Translation is applied to learn how to map the former into the latter. Statistical graphical models are used in [Tellex *et al.*, 2011] to enable a mapping between words and syntactic parse structures with concrete objects, places, paths and events in the real world. The system assigns an interpretation to a sentence by finding the set of groundings that most likely fits the syntactic structure. A beam search is performed across all the entities and concepts described in a semantic map, so that the final interpretation is jointly performed with grounding. In [Matuszek *et al.*, 2012b], a probabilistic CCG is induced to map natural route instructions to robot executable commands. Parameterization is performed with a log-linear model over training data of sentences paired with the corresponding commands, encoded using a specific robot control language. In [Thomason *et al.*, 2015], a weighted CCG is used to parse utterances into λ-calculus expressions, whose variables are grounded by querying a knowledge base of facts about the environment. In addition, the system is able to learn new referring expressions to known objects through dialogs with the user. A similar kind of interaction was already presented in [Kaplan, 2000], where a four-legged robot is instructed through dialog to acquire lexical references for real objects. Once a new object is shown to the robot, the acquiring process is handled by extracting structured visual information from the image and pairing them with the referenced word uttered by the human.

The approach we propose in this work makes use of grounded features extracted from a semantic map about existence of entities in the environment, as well as spatial relations among them, e.g. distance. Such features are used to drive the interpretation process of actions expressed in vocal commands. While some of the aforementioned works did not consider perceptual knowledge as a discriminating factor for the interpretation, other works, such as [Tellex *et al.*, 2011], make a joint use of linguistic and perceptual information. However, in their work perceptual knowledge never modifies syntactic structures that can be wrongly generated from the parser. Similarly to our work, the approach in [Kaplan, 2000; Thomason *et al.*, 2015] deals with the use of words unknown to the robot to refer to objects, even though referents are acquired through dialog. Differently, we make use of a mech-

anism based on Distributional Model of Lexical Semantics [Sahlgren, 2006; Mikolov *et al.*, 2013] together with phonetic similarity functions to achieve robustness (as in [Bastianelli *et al.*, 2015]), while extracting grounded features through the lexical references contained in the semantic map. No further interactions are required, and the acquisition of synonymic expressions for referring to entities is automatically derived by reading large-scale document collections. It is worth noticing that approaches of joint language and perception have been proposed to model the language grounding problem when in presence of grounded attributes, as in [Matuszek *et al.*, 2012a; Krishnamurthy and Kollar, 2013]. Although the underlying idea of these works is similar to ours, our aim is to produce the grounded interpretation at the predicative level, that activates the robotic plan corresponding to the action expressed in an utterance. The findings of such works can be considered as complementary to our proposal.

3 Perceptually-informed Interpretation

In a house servicing robot scenario, learning for grounded *NL interpretation* depends on a wide range of linguistic and perceptual information: the former can be extracted from the utterance transcriptions and refers to the overall linguistic competence of the robot; the latter depends on the map of the currently perceived environment (e.g. the instance `tab1` of the class of `tables`) as this is referenced by linguistic symbols (e.g. *table*). Grounding is thus needed *during* interpretation, in terms of associations between language expressions and elements from the *semantic map*.

Semantic Map. In our setting, a semantic map *SM* is a pair $\langle \mathcal{M}, \mathbf{E} \rangle$, where: $\mathcal{M}$ is a 2D metric map, i.e. the collection of points representing the geometry of the environment as it is directly perceived by visual and depth sensors; $\mathbf{E}$ is the representation of entities existing and perceived in the environment, i.e. recognized by a process of semantic mapping. On entities $\mathbf{e} \in \mathbf{E}$, a function $p(\cdot)$ can be defined in order to refer to their positions made available by $\mathcal{M}$, so that $p(\mathbf{e}) = \langle x_1, y_1 \rangle$ are the coordinates in $\mathcal{M}$ associated with the entity $\mathbf{e}$. Moreover, $\mathcal{LR}(\cdot)$ is an additional function that associates to every $\mathbf{e} \in \mathbf{E}$ the set of its lexical references $w_\mathbf{e}$, that are the expressions used to make linguistic reference to $\mathbf{e}$: it follows that $\mathcal{LR}(\mathbf{e}) = \{w_\mathbf{e} | w_\mathbf{e}$ is a name known to the robot as a reference for $\mathbf{e}\}$. Notice that any $w_\mathbf{e}$ is a name (i.e. a word or a complex expression) given to $\mathbf{e}$ during either a knowledge engineering process or during a Human-Augmented Mapping process. Other attributes for entities $\mathbf{e} \in \mathbf{E}$ may be present, but this information is not relevant here and will be neglected.

Interpretation. In this work, we will rely on the Frame Semantics theory [Fillmore, 1985] to give a linguistic and cognitive basis to the interpretation of the actions encoded in user utterances. Specifically, we will consider the formalization adopted in FrameNet [Baker *et al.*, 1998]. According to such theory, actions, or more generally events, are modeled as *semantic frames*. These are micro-theories about real world situations, e.g. the action of *Taking*. Each frame specifies also the set of participating entities, called *frame elements*, e.g. the THEME representing the object that is moved during the action. For example, for the sentence "*take the book on the table*", a corresponding parsing can be [*take*]$_{\textit{Taking}}$ [*the book on the table*]$_{\text{THEME}}$. Hence, given a sentence s as a sequence of words w_i, i.e. $s = \langle w_1, ..., w_l \rangle$, in our setting an interpretation $\mathcal{I}(s)$ in terms of Frame Semantics determines a set of pairs $\langle \mathtt{f}^i, \mathbf{Arg}^i \rangle$ where $\mathtt{f}^i$ is a frame (each anchored in the text through a *lexical unit* LU, e.g. the verb *take*) and $\mathbf{Arg}^i$ describes the set of arguments of the i-th frame evoked by s. Notice that every $arg^i_j \in \mathbf{Arg}^i$ is a triple $\langle as^i_j, fe^i_j, sh^i_j \rangle$ describing: (i) the span (as^i_j) defined as subsequences of s, so that the span $as^i_j = \langle w_m, ...w_n \rangle$ with $1 \leq m < n \leq l$; (ii) the role label (i.e. frame element, fe^i_j) associated to the spans and drawn from the vocabulary of frame elements $\mathbf{FE}^i$ defined by FrameNet for the current $\mathtt{f}^i$, i.e. $\forall j \ \ fe^i_j \in \mathbf{FE}^i$; and (iii) the semantic head (sh^i_j) of the j-th argument of $\mathtt{f}^i$, i.e. the meaning carrier word w_k of the frame argument. In the above example, where *take* is the LU:

$$\begin{aligned}\mathcal{I}(s) \ = \ & \{\langle Taking, \{ \\ & \langle \langle take \rangle, \text{LU}, take \rangle, \\ & \langle \langle the, book, on, the, table \rangle, \text{THEME}, book \rangle \}\rangle\}\end{aligned}$$

Grounding. In order for the robot to execute the requested command, the corresponding interpretation $\mathcal{I}(s)$ must be grounded. In fact, the semantic frames provided by $\mathcal{I}(s)$ are supposed to trigger grounded command instances as made available by plans (or other behaviors) of the robot. Grounding an instantiated frame in $\mathcal{I}(s)$ requires two steps. First, the frame $\mathtt{f}^i$ in $\mathcal{I}(s)$ must be mapped into a plan: as a consequence, frame arguments must be explicitly associated to their corresponding actors in the plan. Arguments of a plan are directly mapped to frame elements by the so-called *Platform Model*. Once a plan has been selected, its arguments can be paired just with the lexical fillers sh^i_j (e.g. *table*) corresponding to frame elements and these can play the role of anchors for the grounding onto the map: each lexical item can be used to *retrieve* a corresponding instance $\mathbf{e} \in \mathbf{E}$ in the environment (e.g. `tab1`), given the naming associated with it, e.g. $w_\mathbf{e} \in \mathcal{LR}(\mathtt{tab1})$. This lexically-driven grounding is carried out, by applying a lexicalized distance function $g(sh^i_j, w_\mathbf{e})$, that estimates how well the filler sh^i_j matches the entity name $w_\mathbf{e}$. Following [Bastianelli *et al.*, 2015], $g(\cdot, \cdot)$ is estimated as a linear combination between vector descriptions of sh^i_j and $w_\mathbf{e}$, and phonetic similarities. These lexical semantic vectors are acquired through corpus analysis, as in Distributional Lexical Semantic paradigms. They allow to control references to elements modeling synonymy or co-hyponymy, when lexical fillers, such as *photo*, are used to refer to entities with different names, e.g. a `picture`. Phonetic similarities support the interpretation process against possible ASR transcription errors, such as between *pitcher* and *picture*. The maximization of the similarity $g(\cdot, \cdot)$ between fillers and entities corresponds to the minimization of the distance between the corresponding lexical semantic vectors and it can be extensively applied to optimize grounding. Given the set $\mathbf{E}$ of candidate entities in a *SM*, the criterion $\mathcal{OG}$ for grounding frame arguments arg^i_j is defined as follows:

$$\mathcal{OG}(arg^i_j, SM) = \{\mathbf{e} \in \mathbf{E} | \exists w_\mathbf{e} \in \mathcal{LR}(\mathbf{e}), g(sh^i_j, w_\mathbf{e}) > \tau\}$$

where τ is an empirically estimated threshold obeying to application-specific criteria. $g(\cdot,\cdot)$ measures the confidence associated with individual groundings over the relevant lexical vectors. Although different settings of $\mathcal{OG}$ (and therefore of $g(\cdot,\cdot)$) can be designed ([Bastianelli *et al.*, 2015]), this mechanism is extensively used in this paper to locate candidate grounded entities in the *SM* and to code them into perceptual features in the SLU process, hereafter described.

3.1 The Language Understanding Cascade

The proposed interpretation process is based on a cascade of statistical classification steps, modeled as sequence labeling tasks [Croce *et al.*, 2012; Bastianelli *et al.*, 2014a]. The classification is applied to the entire sentence and is modeled as the Markovian formulation of a structured SVM (i.e. *SVM*^{hmm} proposed in [Altun *et al.*, 2003]). In general, this learning algorithm combines a local discriminative model, which estimates the individual observation probabilities of a sequence, with a global generative approach to retrieve the most likely sequence, i.e. tags that better explain the whole sequence. In other words, given an input sequence $\mathbf{x} = (x_1 \dots x_l) \in \mathcal{X}$ of feature vectors $x_1 \dots x_l$, *SVM*^{hmm} learns a model isomorphic to a k-order Hidden Markov Model, to associate $\mathbf{x}$ with a set of labels $\mathbf{y} = (y_1 \dots y_l) \in \mathcal{Y}$.

In this work, a sentence s is intended as a sequence of words w_i, each modeled through a feature vector x_i and associated to dedicated labels y_i, specifically designed for the interpretation process $\mathcal{I}(s)$. Indeed, this process is obtained through the cascade of the Action Detection and Argument Labeling steps, where the latter is further decomposed in the Argument Identification and Argument Classification substeps. Each of these steps is mapped into a different *SVM*^{hmm} sequence labeling task. In the training phase, the SVM algorithm is devoted in associating words to step-specific labels: linear kernel functions are applied to different types of features, ranging from linguistic to perception-based features, and linear combinations of kernels are used to integrate independent properties. At classification time, given a sentence $s = (w_1 \dots w_l)$, the *SVM*^{hmm} efficiently predicts the tag sequence $\mathbf{y}$ using a Viterbi-like decoding algorithm.

The **Action Detection** (AD) step aims at finding all frames evoked by s and filling elements $\mathtt{f}^i$ of the pairs in $\mathcal{I}(s)$. It corresponds to a function $f_{AD}(s, PM, SM)$ as the labeling process depends on linguistic information as well as knowledge derived from the Platform Model (*PM*) and the perceptual knowledge derived from the Semantic Map (*SM*). In our markovian setting, states reflect frame labels, so that the decoding proceeds by detecting lexical units and assigning the proper frame, i.e. an action, in the form of a pair $\langle w_i, \mathtt{f}\rangle$, e.g. *take*-`Taking`. A special *null* label (_) is used to express the status of all other words, e.g. *the*-_ or *book*-_. Each word is represented as a feature vector, defined as follows. Linguistic features here include lexical features (such as the surface or lemma of the current word and its left and right lexical contexts) and grammatical features (e.g. the POS-tag of the current word or the contextual POS-tag n-grams). Information about the robot coming from the *PM* is used to represent actions it is able to perform: these are mapped into frames (e.g. through their corresponding LUs). Given a set of pairing between LUs and frames, for each word in a sentence, boolean features are used to suggest possibly activated frames: in particular, if a word w_i is a verb, then for every frame $\mathtt{f}^j \in F^i$ we set to `true` the corresponding j-th feature for w_i, where F^i is the subset of frames that can be evoked by w_i as in the *PM*. In addition, features derived from the perceptual knowledge are used in the AD step as they are extracted from the *SM*. These "perception-based" features combine the information derived by the lexical grounding function with the syntactic dependency tree associated with s. In particular, given a verb v_i and $n(v_i) = \{w_j | w_j$ is a noun in the dependency (sub)tree rooted in $v_i\}$, the following features are associated to each word $w_j \in n(v_i)$: (i) the number of nouns governed by v_i, i.e. $|n(v_i)|$, (ii) the number of referred entities, i.e. $|\bigcup_{w_j \in n(v_i)} \Gamma(w_j)|$, where[1]:

$$\Gamma(w_j) = \{\mathbf{e} \in \mathbf{E} | \exists w_{\mathbf{e}} \in \mathcal{LR}(\mathbf{e}), g(w_j, w_{\mathbf{e}}) > \tau\}$$

and (iii) the average spatial distance between these entities according to the *SM*.

For each identified frame $\mathtt{f}^i \in \mathcal{I}(s)$, the **Argument Identification** (AI) step detects all its argument spans as^i_j with the corresponding semantic heads sh^i_j. This process starts filling the missing elements of each j-th argument $arg^i_j \in \mathbf{Arg}^i$: more formally for a given frame $\mathtt{f}^i$, with lexical unit LU_i, the AI process can be summarized as the function $f_{AI}(s, SM, \mathtt{f}^i, \text{LU}_i)$. According to the proposed markovian approach, given s and the detected frame $\mathtt{f}^i$, states now denote argument boundaries between individual $arg^i_j \in \mathbf{Arg}^i$ according to the `IOB2` notation: the Begin (`B`), Internal (`I`) or Outer (`O`) tags are assigned to each token. In our running example, the final labeling is represented as `O`-*take* `B`-*the* `I`-*book* `I`-*on* `I`-*the* `I`-*table*. In this step, the same morpho-syntactic features adopted in the AD step are used together with the frame $\mathtt{f}^i$ under analysis. Moreover, dedicated features derived from the perceptual knowledge are introduced: a boolean feature is set to `true` for all and only the nouns w_j such that $\Gamma(w_j) \neq \emptyset$, with $\Gamma(w_j)$ containing candidate entities referred by w_j, as above; for prepositions p_j, given their syntactic dependent w^{dep}_j, a second boolean feature is set to `true` iff $\Gamma(w^{dep}_j) \neq \emptyset$. The number of nouns on the left and on the right of p_j are also used as features in its corresponding feature vector. Finally, for each preposition p_j, we also retrieve its syntactic governor in the tree w^{gov}_i and measure the average Euclidean distance in *SM* between entities in $\Gamma(w^{dep}_j) \cup \Gamma(w^{gov}_i)$: if this score is under a given threshold, a spatial feature is set to `near`, replacing the default value of `far`.

In the **Argument Classification** (AC) step, given the frame $\langle \mathtt{f}^i, \mathbf{Arg}^i \rangle \in \mathcal{I}(s)$, each $arg^i_j \in \mathbf{Arg}^i$ is labeled through its frame element fe^i_j, e.g. THEME to the argument *the book on the table*. In this step states correspond to role labels. Classification here exploits only linguistic features, as grounded information extracted from *SM* is not essential in this subtask. Morpho-syntactic features are extracted from s, and semantic features, such as $\mathtt{f}^i$ and `IOB2` tags coming from the

[1] Γ also depends on the *SM* but it is omitted for simplicity.

previous stages (AD and AI), are considered. In addition, a Distributional Model (*DM*) of Lexical Semantics is applied to generalize the argument semantic head sh^i_j: the distributional (vector) representation for sh^i_j is thus introduced to extend the feature vector corresponding to each $w \in as^i_j$. Given a frame $\mathtt{f}^i$, the set of $as^i_j \in \mathbf{Arg}^i$, the AC function can thus be written as: $f_{AC}(s, \mathtt{f}^i, \text{Lu}_i, \mathbf{Arg}^i, DM)$.

4 Experimental Evaluation

The contribution of perceptual information has been evaluated in the semantic interpretation of utterances in a house Service Robotics scenario. The evaluation is carried out using the Human-Robot Interaction Corpus (HuRIC, [Bastianelli *et al.*, 2014b]) a collection of utterances annotated with semantic predicates and paired with (possibly multiple) audio files. Utterances are annotated with linguistic information of various kinds (from morpho-syntax to semantic frames). HuRIC contains 860 audio files for 527 sentences.

Since linguistic information was provided in HuRIC without an explicit representation of the environment, we extended the corpus by pairing each utterance with a possible reference environment: each s in HuRIC is paired with a generated semantic map (*SM*) reflecting the disposition of entities matching the interpretation, so that perceptual features can be consistently derived for each s. Extended examples are of the form $\langle s, SM \rangle$. The map generation process has been designed to reflect real application conditions. First, we built a reference knowledge base (KB) acting as domain model and containing classes that describe the entities of a generic home environment. Then, for each sentence s, the corresponding *SM* is populated with the set of referred entities, plus a control set of 20 randomly-generated additional objects, all taken from the KB. The naming function $\mathcal{LR}$ has been defined simulating the lexical references introduced by a process of Human-Augmented Mapping. The set of possible lexical alternatives (from which such $\mathcal{LR}$ draws) has been designed to simulate free lexicalization of entities in the *SM*. For every class name in the KB, a range of possible polysemic variations has been defined, by automatically exploiting lexical resources, such as WordNet [Miller, 1995], or by corpus-analysis. The final set has been then validated by human annotators. As an example the class `table` is referred through the following variations: *table*, *desk* and *board*. The above lexical variation allows augmenting the data set as each training sentence can be paired with more than one *SM*. For each sentence s, two *SM*s have been generated: in one map each entity (e.g. `table`) referred in s has the corresponding class name (e.g. *table*), and another in which all entities are named through lexical variations (e.g. *desk*). In this way, 1,054 examples have been generated. The distributional analysis underlying $g(\cdot, \cdot)$ and the *DM* vectors has been acquired according to a Skip-gram model [Mikolov *et al.*, 2013], through the `word2vec` tool. By applying the settings *min-count=50*, *window=5*, *iter=10* and *negative=10* onto the UkWaC corpus we derived 250 dimensional word vectors for more than $110,000$ words. The *SVM*^{hmm} algorithm has been implemented within the KeLP framework [Filice *et al.*, 2015].

Table 1: Results w.r.t. the Action Detection step

	P	**R**	**F**	**RER**
Baseline	79.79%	68.56%	73.75%	-
noPM/noSM	93.80%	95.56%	94.67%	0.0%
onlyPM	94.61%	95.57%	95.09%	7.9%
PerfG	96.25%	96.42%	96.33%	31.1%
LexG	95.82%	95.99%	**95.91**%	23.3%

Table 2: Results w.r.t. the Argument Identification step

	P	**R**	**F**	**RER**
Baseline	75.46%	93.21%	83,40%	-
noPM/noSM	88.99%	92.56%	90.74%	0.0%
PerfG	94.48%	94.75%	94.62%	41.9%
LexG	94.02%	94.56%	**94.29**%	38.3%

Table 3: Results w.r.t. the Argument Classification step

	P	**R**	**F**	**RER**
Baseline	21.78%	21.78%	21.78%	-
noDM	94.93%	94.93%	94.93%	0.0%
DM	95.31%	95.31%	**95.31**%	15.8%

Measures have been carried out on four tasks, all according to a 5-fold evaluation schema. The first three correspond to the individual interpretation steps, namely the AD, AI and AC. In these tests, we assume that the input information of the task corresponds to the gold annotation even if it depends from a previous processing step. For each run a baseline has been estimated to determine the task complexity when minimal information is considered. The last test concerns the analysis of the end-to-end interpretation chain. It thus corresponds to the ability of translating a vocal command into a fully grounded and executable command.

The tasks in which perceptual knowledge is involved are the AD and AI. We considered several settings: (i) no perceptual information is considered (noPM/noSM); (ii) perfect grounding information is assumed, that is gold information about entities is provided instead of using any $\mathcal{OG}$ function over the semantic map (PerfG); (iii) grounding information is based on the $\mathcal{OG}$ set, built by the $\mathcal{LR}$ function introduced before (LexG). An additional run has been carried out for AD, by considering only the Platform Model as the source of self-knowledge (onlyPM). Results obtained in every run are reported in terms of Precision (P), Recall (R) and F-Measure (F) as a micro-statistics across the 5 folds. The contribution of perceptual information is emphasized in terms of Relative Error Reduction (RER) over F-measure w.r.t. the system setting relying just on linguistic information (noPM/noSM).

Action Detection. The results about the AD step are reported in Table 1. The set of frames involved by some HuRIC sentence include 17 frames, with an average ambiguity of ~ 1.21 per lexical unit, considering that some LUs may evoke different frames (such as *take* vs. *Taking* or *Bringing*). The baseline is obtained by choosing a frame randomly selected among the possible ones suggested by each LU occurring in a sentence. As shown in Table 1, the proposed labeling outperforms the baseline even when only linguistic information is employed (noPM/noSM). Further improvement is achieved when perception comes into play, reaching a significant 95.65% of F in the LexG setting. It is noticeable

Table 4: Results of the whole interpretation chain, w.r.t. three testing scenarios

				Gold transcr., Gold Info.			Gold transcr., CoreNLP			ASR, CoreNLP		
AD	AI	AC	Ground.	P	R	F	P	R	F	P	R	F
noPM/noSM	noPM/noSM	noDM	$Ident$	43.7%	44.9%	44.3%	42.0%	41.3%	41.7%	32.9%	27.3%	29.8%
LexG	LexG	DM	$\mathcal{OG}_{max}$	74.5%	74.7%	**74.6%**	68.9%	65.8%	**67.3%**	59.3%	46.5%	**52.1%**

that the RER, whose upper bound is 31.1% as for the gold grounding condition (PerfG), reaches 23.3% when the proposed operational grounding method $\mathcal{OG}$ is employed. This allows to recover from misclassifications of the noSM/noPM setting, as in "*take the mobile into the bedroom*", for which the wrong TAKING frame is corrected to BRINGING when perceptual evidences about the distance between the *mobile* and the *bedroom* are made available.

Argument Identification. The results are reported in Table 2. The baseline for the AI task exploits a purely syntactic approach: given a frame and its corresponding lexical unit LU, one argument for each branch of the dependency tree rooted at a LU is assumed. Coordination structures have been skipped. As it is clear from Table 2, perceptual knowledge plays a crucial role for this task, as shown by the difference of $\sim$4 absolute points between the other settings and the noPM/noSM one. Lexical grounding features help substantially in locating argumental chunks of a sentence, especially in ambiguous structures., e.g. left prepositional attachments. For example, the fragment *the book on the table* may correspond to one single argument in which *on the table* is a spatial modifier of *the book*. But when no book is on any table in the environment, it is more likely to correspond to two arguments, i.e. THEME for *the book* and the destination role (GOAL) for *on the table*.

Argument Classification. The AD task has been tested across two runs: in the first one (DM) the word embeddings computed over the UkWaC corpus is considered while in the second one (noDM) it is neglected. The AC baseline assigns randomly, to each argument detected in the AI stage, one frame element, among those used at least once in HuRIC for that frame. Beside outperforming the baseline, further improvements are achieved when distributional information about words is adopted. The *DM* injects beneficial lexical generalization into training data: frame element of arguments whose semantic heads are close in the vector space are seemingly tagged. For example, if *the book* in the training sentence "*take the book*" is the THEME of a TAKING frame, similar arguments for the same frame will receive the same role label as *notepad* in "*grab the notepad*".

Evaluating the whole interpretation chain. The last experiment aims at measuring the performances of an end-to-end cascade involving all the three AD, AI and AC steps. We also performed a fourth grounding step that converts instantiated semantic frames into executable commands, as discussed in Section 3. As a consequence, the performances here are evaluated in terms of P, R and F over the fully grounded command, i.e. an action correctly specified together with all its arguments. Two runs have been performed, whose results are reported in Table 4. In the first run (the first row), neither perception (noPM/noSM) nor any lexical smoothing (noDM) are considered in the AD, AI and AC steps. In addition, *lexical identity* (*Ident*) is also used to map the arguments of the plan to entities of the semantic map. In a second run (second row), information coming from the semantic map (LexG) and lexical smoothing (DM) are both applied, and a $\mathcal{OG}_{max}$ function is used to ground the arguments of the plan, considering the sole entity maximizing the $g(\cdot,\cdot)$ function among the candidates. For every run, the system has been tested in three scenarios. The first scenario reflects optimal conditions for speech recognition and morpho-syntactic parsing and the chain is fed with perfect transcriptions and gold morpho-syntactic information. In a second more realistic scenario, morpho-syntactic parsing is carried out by the CoreNLP parser over correct transcriptions. Finally, the last scenario reflects a real voice interaction where audio files of the commands are processed by an off-the-shelf free-form ASR engine, and the best transcription hypothesis is provided to the interpretation chain. These last two settings reproduce realistic working conditions, where intermediate errors can be introduced. For completeness, the performances of the CoreNLP and the adopted speech recognition engine are reported below. The POS-tagging accuracy is of 93.98%, while dependency parsing has an Unlabeled and Labeled Attachment Score of 87.57% and of 85.20% respectively. Speech recognition scored 79.82% of Precision@1 over the audio commands from HuRIC. When perceptual knowledge from the Semantic maps, Distributional Models and robust grounding techniques are employed, the performances significantly grow. This drop is also due to the difficulty in grounding plan arguments over maps generated with lexical variability. The average gain of $\sim$26 points of F1 testify the successful application of our approach also in real application conditions.

5 Conclusions and Future Work

In service robotics, Spoken Language Understanding must specifically model grounded NL interpretation dealing with physical world and perceptual knowledge. In this paper, we present a discriminative approach to SLU for Interactive Robotics where, in addition to conventional linguistic features, perceptual information is made available in the form of a semantic map. Our approach has three advantages: (i) we make the language understanding process sensitive to grounded reasoning, (ii) we improve interpretation accuracy (up to 38% RER) with different potential interpretations against different perceived environments and (iii) the output representation of an interpreted command is already augmented through the grounding of all linguistic elements. The model proposed here still relies on discretized representations of perceptual knowledge such as semantic maps. However, an effort for directly integrating information coming from other perceptual subsystems, e.g. vision, is enabled and could represent the next step of this work.

References

[Altun *et al.*, 2003] Yasemin Altun, I. Tsochantaridis, and T. Hofmann. Hidden Markov support vector machines. In *Proc. of ICML*, 2003.

[Baker *et al.*, 1998] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley framenet project. In *Proceedings of ACL and COLING*, pages 86–90, 1998.

[Bastianelli *et al.*, 2013] Emanuele Bastianelli, Domenico Daniele Bloisi, Roberto Capobianco, Fabrizio Cossu, Guglielmo Gemignani, Luca Iocchi, and Daniele Nardi. On-line semantic mapping. In *Advanced Robotics (ICAR), 2013 16th International Conference on*, pages 1–6, Nov 2013.

[Bastianelli *et al.*, 2014a] Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, Roberto Basili, and Daniele Nardi. Effective and robust natural language understanding for human-robot interaction. In *Proceedings of ECAI 2014*. IOS Press, 2014.

[Bastianelli *et al.*, 2014b] Emanuele Bastianelli, Giuseppe Castellucci, Danilo Croce, Roberto Basili, and Daniele Nardi. Huric: a human robot interaction corpus. In *Proceedings of LREC 2014*, Reykjavik, Iceland, may 2014.

[Bastianelli *et al.*, 2015] Emanuele Bastianelli, Danilo Croce, Roberto Basili, and Daniele Nardi. Using semantic models for robust natural language human robot interaction. In *AI* IA 2015, Advances in Artificial Intelligence*, pages 343–356. Springer International Publishing, 2015.

[Bos and Oka, 2007] Johan Bos and Tetsushi Oka. A spoken language interface with a mobile robot. *Artificial Life and Robotics*, 11(1):42–47, 2007.

[Chen and Mooney, 2011] David L. Chen and Raymond J. Mooney. Learning to interpret natural language navigation instructions from observations. In *Proceedings of the 25th AAAI Conference on AI*, pages 859–865, 2011.

[Croce *et al.*, 2012] D. Croce, G. Castellucci, and E. Bastianelli. Structured learning for semantic role labeling. *Intelligenza Artificiale*, 6(2):163–176, 2012.

[Diosi *et al.*, 2005] Albert Diosi, Geoffrey R. Taylor, and Lindsay Kleeman. Interactive SLAM using laser and advanced sonar. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation, ICRA 2005, April 18-22, 2005, Barcelona, Spain*, pages 1103–1108, 2005.

[Filice *et al.*, 2015] Simone Filice, Giuseppe Castellucci, Danilo Croce, and Roberto Basili. Kelp: a kernel-based learning platform for natural language processing. In *Proceedings of ACL2015: System Demonstrations*, Beijing, China, July 2015.

[Fillmore, 1985] Charles J. Fillmore. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254, 1985.

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

[Kaplan, 2000] F. Kaplan. Talking AIBO: First experimentation of verbal interactions with an autonomous four-legged robot. In *Proceedings of the CELE-Twente workshop on interacting agents*, 2000.

[Kollar *et al.*, 2010] Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. Toward understanding natural language directions. In *Proceedings of the 5th ACM/IEEE*, HRI '10, pages 259–266, Piscataway, NJ, USA, 2010.

[Krishnamurthy and Kollar, 2013] Jayant Krishnamurthy and Thomas Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *TACL*, 1:193–206, 2013.

[Kruijff *et al.*, 2007] Geert-Jan M. Kruijff, H. Zender, P. Jensfelt, and Henrik I. Christensen. Situated dialogue and spatial organization: What, where... and why? *International Journal of Advanced Robotic Systems*, 4(2), 2007.

[Matuszek *et al.*, 2012a] Cynthia Matuszek, Nicholas FitzGerald, Luke S. Zettlemoyer, Liefeng Bo, and Dieter Fox. A joint model of language and perception for grounded attribute learning. In *ICML*. icml.cc / Omnipress, 2012.

[Matuszek *et al.*, 2012b] Cynthia Matuszek, Evan Herbst, Luke S. Zettlemoyer, and Dieter Fox. Learning to parse natural language commands to a robot control system. In Jaydev P. Desai, Gregory Dudek, Oussama Khatib, and Vijay Kumar, editors, *ISER*, volume 88 of *Springer Tracts in Advanced Robotics*, pages 403–415. Springer, 2012.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[Miller, 1995] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[Nüchter and Hertzberg, 2008] Andreas Nüchter and Joachim Hertzberg. Towards semantic maps for mobile robots. *Robot. Auton. Syst.*, 56(11):915–926, 2008.

[Sahlgren, 2006] Magnus Sahlgren. *The Word-Space Model*. PhD thesis, Stockholm University, 2006.

[Tanenhaus *et al.*, 1995] M. Tanenhaus, M. Spivey-Knowlton, K. Eberhard, and J. Sedivy. Integration of visual and linguistic information during spoken language comprehension. *Science*, 268:1632–1634, 1995.

[Tellex *et al.*, 2011] S. Tellex, T. Kollar, S. Dickerson, M.R. Walter, A.G. Banerjee, S. Teller, and N. Roy. Approaching the symbol grounding problem with probabilistic graphical models. *AI Magazine*, 32(4), 2011.

[Thomason *et al.*, 2015] Jesse Thomason, Shiqi Zhang, Raymond Mooney, and Peter Stone. Learning to interpret natural language commands through human-robot dialog. In *Proceedings of the 2015 International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1923–1929, Buenos Aires, Argentina, July 2015.