Talking Robots

Emanuele Bastianelli

Ph.D. student at Department of Civil Engineering and Computer Science Engineering
Tor Vergata University of Rome, Italy
Research associate at Department of Computer, Control, Management Engineering
Sapienza University of Rome, Italy
bastianelli@ing.uniroma2.it

Abstract. In the last years robotic platforms have appeared in many research and everyday life contexts. An easy way of interacting with them has then become a necessity. *Human Robot Interaction* is the research field that aims at studying how robots can interact with humans in the most natural way. In this work we will present preliminary studies that we have been done in this direction, focusing on *Natural Language* based interaction, with particular attention to the *grounding* problem. In particular, we will study how *Statistical Machine Learning* techniques can be applied to Natural Language as it is used to interact with robots. Moreover, we will also investigate how this approach can be integrated in such complex systems.

1 Introduction

Robots are slowly becoming part of everyday life, as they are being marketed for commercial applications (viz. telepresence, cleaning or entertainment). As a consequence, the ability to interact with a non-expert user is becoming a key requirement. The *Human Robot Interaction* (HRI) field aims at realizing robotic systems that offer a level of interaction the more natural as possible. This means providing robots with sensory systems capable of understanding and replicating human languages, such as speech, gestures, voice intonation, pragmatic interpretation, and any other non-verbal interaction. The ultimate goal of this research area is to provide robots with the ability of solving human languages references in the application context they belong, such as the real world (e.g. assigning the right coordinates to the phrase *the kitchen*) or an abstract world (e.g. solving anaphoric references in the domain of the discourse). This cognitive process, that is natural and implicit among human beings, is commonly called *grounding*.

Our research will investigate the problems related to the natural language analysis involved in the design of HRI systems. To this aim, we will explore the possibility of reusing approaches that have been largely applied in different *Natural Language Understanding* (NLU) tasks and testing their applicability in the HRI field. In particular, we will focus on finding a bridge between the linguistic knowledge expressed in spoken commands and the robot representation of the world as a support for the grounding process.

2 Motivations and Background Works

Among the different kinds of interaction treated by HRI, we will focus on those aspects involving natural language. User utterances can be recognized and transcribed by *Automatic Speech Recognition* systems (ASR) that in the last years have become more and more accessible and powerful. The main issue is that in order to translate user utterances into robotic actions, we need to understand their meaning. For instance, from the sentence "*take the bottle on the table*", we need to provide the command corresponding to the action of *taking* . Moreover, we need to identify the relation holding between *the bottle* and *the table.* This semantic information can be crucial as well to ground linguistic expressions into objects as they are represented in the robot *set of beliefs* (i.e. robot knowledge and perceptions).

To fill the gap between the robot world representation and the linguistic knowledge expressed in user utterances, we need to extract the meaning from a sentence and represent it in a suitable form. Grammar-based ASR systems often offer the possibility to attach semantic primitives to each grammar rule. The meaning representation is obtained as the composition of all the primitives explored during the decoding. Such approach has been largely adopted in the robotic field, as in [4]. Grammars indeed have the limit of covering just a segment of the language. If we want to realize more general HRI systems, and thus to cover a wider range of linguistic phenomena, we need to rely on free-form speech SR engines. Unfortunately, this kind of systems do not provide any kind of additional information, besides the plain transcription of utterances. A representation of their meaning can be only obtained by an external semantic parsing process. *Natural Language Processing* (NLP) approaches based on formal languages have found wide application in the HRI field, e.g. semantic parsing with *Combinatory Cathegorial Grammars* (CCG), as in [5], where a way of obtaining a meaning representation based on *Discourse Representation Structures* [9] directly from the speech recognition is used. Similarly, in [17], CCGs are used to produce a representation in term of *Hybrid Logics Dependency Semantics* [16] logic form. However, the overall attention has recently shifted towards the application of *Statistical Learning* techniques, reflecting the will of designing more general solutions. Several fields of research have shown a growing interest in HRI, giving the chance to apply these techniques in this area. Experts with different backgrounds proposed their own approach, mainly coming from Robotics, Computational Linguistics and Cognitivism.

The problem of grounding natural language symbols into robot representations of the world has been mostly explored in developing system for tasks as Human Augmented Mapping or able to follow route instructions. In [18], a simulated robot system called MARCO able to follow route instruction in a virtual environment is presented. Here spoken commands are parsed using *compound action specifications* to model which actions to take under which conditions. These structures capture the commands in route instructions by modeling the surface meaning of a sentence as a verb-argument structure, and are obtained after a natural language processing chain. This work has been continued and extended in [6], where Statistical Learning has been applied to learn how to map commands in the corresponding logical form-like structure. This represent the robot instruction that can be directly executed and implicitly resolves the grounding of all the entities. The work in [22] proposes a system that learns to follow navigational natural language directions by apprenticeship from routes through a map paired with English descriptions. Reinforcement learning algorithm is applied to determine what portions of the language describe which aspects of the route.

Other works have been inspired by novel spatial semantic theories. In [14] the problem of executing natural language directions is formalized through *Spatial Description Clauses* (SDCs), a structure that can hold the result of the spatial semantic parsing in terms of spatial roles. The same representation has been exploited in a subsequent work [21], where the probabilistic graphical model theory is applied to parse each instruction into a structure called *Generalized Grounding Graph* (G^3). Here the SDCs are used as a base component of the more general structure of the G^3, that represents both semantic and syntactic information of the spoken sentence. In some cases, the construction of the representation is taken into account as in [11], where the robot learns the features of the environment, through the use of narrated guided tours. In this work, the robot builds both the metrical and topological representations of the environment during the tour. Spatial and semantic information are then associated to the evolving situations through *events labeling*, that occur during a tour, and are later attached to the nodes of a topological graph.

However, the approaches proposed so far have only taken into account single aspects (e.g. deep analysis of solving spatial relations [14, 21]) of the overall linguistic analysis necessary to realize a complete grounding process. The complexity of the problem is higher and is well described in [20]. Here, it is stated that a complete natural language HRI system should be able to: (*i*) react in the same time frame of a human; (*ii*) process all stages of language processing in a concurrent way; (*iii*) own the capability of understanding *spoken* language; (*iv*) decode multi-modal cues, such as linguistic expressions accompanied by gestures; (*v*) share the perspective on the world and on events with its interlocutors; (*vi*) start interaction to support bidirectional communications. All these features can constrain possible interpretations of the language, biasing the grounding process. It arises that the level of natural language analysis therefore needed is high and complex, as different informations, corresponding to different levels of semantics, need to be extracted and provided to the system. Moreover, this results in a sophisticated interaction schema among the system modules (e.g. NLU processors, inference engines over knowledge bases, perception systems).

The need to re-elaborate the problem from this point of view is being perceived by the community. Complex architectures have been already realized for tasks such as *Question Answering*, where the cooperation of structured NLP modules and other processors is fundamental. In order to maximize the replicability and adaptability, we argue that similar approaches should be followed in the implementation of HRI interfaces. One of our purposes is to study the applicability of robust NLP techniques that have been already adopted for other tasks.

Following this direction, as a basic step of our research we contributed to the development of a prototype robot for *Human-Augmented Mapping* that is being used for experimental purpose. The data gathered during the experiments will be used in this research. In the meanwhile, a corpus of spoken commands is being collected using a web interface. It contains audio files paired with the corresponding transcriptions. Each transcription is annotated according to different semantic formalisms, describing the linguistic knowledge we want to capture. This corpus should become a useful resource for several tasks, e.g. training specific learning algorithms.

3 Theories and Methods

An hypothetical NLU processing chain of a HRI system has to deal with audio processing and transcription, meaning understanding and dialogue management. The first module consists in ASR engine. To improve the grounding process, the module can be extended with the capability of detecting the source of the speech. This could assist, in fact, the reference point identification of certain spatial expressions (e.g. "*the door on my right*"). Morphological analysis and syntactic parsing are performed during the second step, as they can add crucial information for further semantic processing. This latter is the core of the NLU chain. Different semantic parsers can be used in parallel or in cascade, as the information generated by one such parser can be useful for others. During this step, the modules can also require an interaction with external resources, such as Linguistic Thesaurus or Knowledge Bases. This might be useful in discarding unlikely interpretations and consequently leading the system to consider other hypotheses from the ASR. Finally, the utterances should be enriched with all the meaning representations needed to correctly ground it in the robot set of beliefs. Dialogical interaction can be managed by a dialogue system that interacts with each step of the process.

In our research, we will mainly focus on the semantic analysis part of the chain. In fact, while robust tools for ASR (e.g. Microsoft Speech Platform, Google Speech API or CMUSphix [23]) and for morpho-syntactic analysis (e.g. Stanford CoreNLP [13]) are available, semantic parsing must be designed from scratch. Although semantic processors exist, they are not always free and, more important, they offer just one level of semantic analysis. As stated in Section 2, we need different levels of information; consequently, our HRI system should rely on several semantic parsers. We need then to define which are the aspects of the world we want to model through semantic analysis. First, in order to be useful a robot is expected to perform the actions corresponding to the received commands; second, these actions take place in an physical environment. Looking back at linguistic theories that studied how these two aspects are conveyed through linguistic knowledge, we found that *Frame Semantics* [10] and *Holistic Spatial Semantics* [24] offer models of interpretation suitable for our purpose. The first generalizes actions or, more generally, experiences representing them as *Semantic Frames*. Each frame describes a scene or the general concept behind an action, enriched by a set of *semantic arguments* that play specific roles with respect to the frame. Robot actions can then be linked with the semantic frame corresponding to that action. For example, in the sentence "*take the book on the table*", the semantic frame related to the action of *Taking* is evoked by the verb *take*. The semantic role THEME (i.e. the entity taken during the *Taking* action) is here expressed and represented by *the book on the table*. Similarly, Holistic Spatial Semantics explains the spatial referring expressions contained in sentences in terms of spatial relations composed by spatial roles. Considering the previous example, the words *book* and *table* are related through the preposition *on*, that holds the spatial relation and plays the role of SPATIAL_INDICATOR, while the other two are respectively the TRAJECTOR and the LANDMARK. These two representations can collaborate to model the sentence meaning in a complete way. One more issue to be addressed is that these representations are not designed to work together, so further research about a formalism that should act as a general-purpose semantic container representation need to be done.

In the first step of this research many of the aspects so far reported have been individually examined, and solution based on novel NLP techniques have been proposed. In [2] we propose a re-ranking approach to get the best speech transcription from the set of different hypotheses produced by an ASR system. The ranking function is learned through a *Support Vector Machine* (*SVM*) exploiting a combination of different kernels capturing syntactic and semantic aspects of the utterances (e.g. *Smoothed Partial Tree Kernels* [8]). Moreover, the linguistic problem of extracting semantic representations from natural language expressions has been proposed in tasks as *Semantic Role Labeling* (SRL)[19] and *Spatial Role Labeling* (SpRL)[15]. We developed SRL [7] and SpRL [3] systems that model the problem as a sequential labeling task, exploiting specific formulations of SVMs, as $SVM^{Multiclass}$ [12] and SVM^{Hmm} [1]. These systems have not yet been used together. Their application in a HRI architecture, using the robotic prototype we developed, deserves further investigation.

Another aspect our research wants to explore is the use of dialogical mechanisms to improve the grounding process. Getting the meaning of a sentence may be insufficient to correctly ground the linguistic references. In fact, these can refer to objects or positions in the real world as well as entities in the abstract domain of the discourse (e.g. anaphoric references). Providing the robot with a more complex level of interaction, such as the ability to ask for clarifications about ambiguous expressions, can improve the grounding capability of the robot. Similarly, the system could use dialogue to learn user-specific linguistic references, such as new terms or particular ways of calling objects, or new syntactic forms. This dialogue would be exploited to update the general knowledge of the robot, adding new concepts in a knowledge base or feeding the ASR grammars with new rules. In our robotic platform, we started modeling the dialogue with *Petri-Net Plans*. They can drive the overall behavior of the robot, by managing the interaction among all the modules, including the NLU chain. The integrated representation of dialogue and robot actions is another issue that we intend to address in our research.

We are aware that this proposal is a starting point for an analysis that will be wider and longer. Among all those aspects that should be examined and modeled in a HRI system, we took into account only the two (i.e. actions and spatial references) we considered fundamental to our ends. Future researches might investigate the study of temporal relations expressed in natural language. In parallel, we want to investigate and foster the reuse of robust NLP solutions in a field where single aspects of the problem have been explored, without converging to a common point.

References

1. Altun, Y., Tsochantaridis, I., Hofmann, T.: Hidden Markov support vector machines. In: Proceedings of the ICML (2003)
2. Basili, R., Bastianelli, E., Castellucci, G., Nardi, D., Perera, V.: Kernel-based discriminative re-ranking for spoken command understanding in hri. In: Proceedings of Ai*iA '13. p. to appear (2013)
3. Bastianelli, E., Croce, D., Nardi, D., Basili, R.: Unitor-hmm-tk: Structured kernel-based learning for spatial role labeling. In: Proceedings of SemEval-2013. Atlanta, Georgia, USA (June 2013)
4. Bos, J.: Compilation of unification grammars with compositional semantics to speech recognition packages. In: COLING (2002), `http://dblp.uni-trier.de/db/conf/coling/coling2002.html#Bos02`
5. Bos, J., Oka, T.: A spoken language interface with a mobile robot. Artificial Life and Robotics 11(1), 42–47 (2007)
6. Chen, D.L., Mooney, R.J.: Learning to interpret natural language navigation instructions from observations. In: Proceedings of AAAI '11. pp. 859–865 (2011)
7. Croce, D., Castellucci, G., Bastianelli, E.: Structured learning for semantic role labeling. Intelligenza Artificiale 6(2), 163–176 (2012)
8. Croce, D., Moschitti, A., Basili, R.: Structured lexical similarity via convolution kernels on dependency trees. In: EMNLP. pp. 1034–1046 (2011)
9. Curran, J., Clark, S., Bos, J.: Linguistically motivated large-scale nlp with c&c and boxer. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions. pp. 33–36. Association for Computational Linguistics, Prague, Czech Republic (June 2007), `http://www.aclweb.org/anthology/P07-2009`
10. Fillmore, C.J.: Frames and the semantics of understanding. Quaderni di Semantica 6(2), 222–254 (1985)
11. Hemachandra, S., Kollar, T., Roy, N., Teller, S.: Following and interpreting narrated guided tours. In: Proceedings of ICRA '11. Shanghai, China (2011)
12. Joachims, T., Finley, T., Yu, C.N.: Cutting-plane training of structural SVMs. Machine Learning 77(1), 27–59 (2009)
13. Klein, D., Manning, C.D.: Accurate unlexicalized parsing. In: Proceedings of ACL'03. pp. 423–430 (2003)
14. Kollar, T., Tellex, S., Roy, D., Roy, N.: Toward understanding natural language directions. In: Proceedings of the 5th ACM/IEEE. pp. 259–266. HRI '10, IEEE Press, Piscataway, NJ, USA (2010)

15. Kordjamshidi, P., Van Otterlo, M., Moens, M.F.: Spatial role labeling: Towards extraction of spatial relations from natural language. ACM Trans. Speech Lang. Process. 8(3), 4:1–4:36 (Dec 2011), http://doi.acm.org/10.1145/2050104.2050105
16. Kruijff, G.J.M.: A Categorial-Modal Logical Architecture of Informativity: Dependency Grammar Logic & Information Structure. Ph.D. thesis, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic (April 2001)
17. Kruijff, G.J.M., Zender, H., Jensfelt, P., Christensen, H.I.: Situated dialogue and spatial organization: What, where…and why? International Journal of Advanced Robotic Systems 4(2) (2007)
18. MacMahon, M., Stankiewicz, B., Kuipers, B.: Walk the talk: connecting language, knowledge, and action in route instructions. In: Proceedings of AAAI '06. pp. 1475–1482. AAAI Press (2006)
19. Palmer, M., Gildea, D., Xue, N.: Semantic Role Labeling. Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers (2010)
20. Scheutz, M., Cantrell, R., Schemerhorn, P.: Toward humanlike task-based dialogue processing for human robot interaction. AI Magazine 34(4), 64–76 (2011)
21. Tellex, S., Kollar, T., Dickerson, S., Walter, M.R., Banerjee, A.G., Teller, S., Roy, N.: Approaching the symbol grounding problem with probabilistic graphical models. AI Magazine 34(4), 64–76 (2011)
22. Vogel, A., Jurafsky, D.: Learning to follow navigational directions. In: Proceedings of ACL '10. pp. 806–814. Association for Computational Linguistics, Stroudsburg, PA, USA (2010)
23. Walker, W., Lamere, P., Kwok, P., Raj, B., Singh, R., Gouvea, E., Wolf, P., Woelfel, J.: Sphinx-4: A flexible open source framework for speech recognition (2004)
24. Zlatev, J.: Spatial semantics. Handbook of Cognitive Linguistics pp. 318–350 (2007)