Textual Inference and Meaning Representation in Human Robot Interaction

Bastianelli Emanuele[1], **Giuseppe Castellucci**[2], **Danilo Croce**[3], **Roberto Basili**[3]
[1]DICII, [2]DIE, [3]DII
University of Roma, Tor Vergata
00133 Roma, Italy
{bastianelli,castellucci}@ing.uniroma2.it
{croce,basili}@info.uniroma2.it

Abstract

This paper provides a first investigation over existing textual inference paradigms in order to propose a generic framework able to capture major semantic aspects in Human Robot Interaction (HRI). We investigate the use of general semantic paradigms used in Natural Language Understanding (NLU) tasks, such as Semantic Role Labeling, over typical robot commands. The semantic information obtained is then represented under the *Abstract Meaning Representation*. AMR is a general representation language useful to express different level of semantic information without a strong dependence to the syntactic structure of an underlying sentence. The final aim of this work is to find an effective synergy between HRI and NLU.

1 Introduction

As robots are being marketed for consumer applications (viz. telepresence, cleaning or entertainment) natural language interaction is expected to make them more appealing and accessible to the end user. The latest technologies in speech recognition are available on cheap computing devices, thus enabling different levels of interaction. The first level needed in HRI is the command understanding. This is a challenging task as it consists not only in understanding the utterance meaning, but also in translating it into the robot-specific command. In the recent years, works about the interpretation of natural language (NL) instructions in a specific environments, e.g. allowing a simulated robot to navigate to a specified location, has been oriented to cover a specific subset of the language (Kruijff et al., 2007; Bos and Oka, 2007). This led to very powerful and formalized systems that are, at the same time, very specific and limited in terms of expressiveness. In many NLP tasks where robustness is crucial, e.g. Question Answering as discussed in (Ferrucci et al., 2010), methods based on Statistical Learning (SL) theory have been used to overcome such issues in the support of complex Textual Inference tasks, as in (Chen and Mooney, 2011).

In this paper, instead of focusing on specific language understanding algorithms, we investigate the combination of state-of-the-art textual inference technologies in order to design effective systems for HRI. The final aim of this research is to propose a unifying framework able to capture semantic aspects as these are needed in the HRI area. We foster the idea that many problems tackled and solved in Natural Language Processing, e.g. Semantic Role Labeling (SRL) (Palmer et al., 2010), can be taken into account for HRI. Existing techniques can be used to automatically acquire useful semantic representations to interpret robot commands as investigated in (Thomas and Jenkins, 2012). Let us consider a domotic scenario where a robot receives vocal instructions, e.g. "*take the book on the table*". We think that the command targeted by this utterance can be expressed through the adoption of semantic roles as defined in existing lexical theories, as discussed in (Fillmore, 1985) or (Levin, 1993). Moreover, the generalization level offered by this representation can be improved to better reflect human instructions with the environment where the robots are acting into. For example, we can extend the semantic roles in order to properly capture spatial as well as temporal expressions. These can be crucial for the robot to understand spatial relations between objects in the space or temporal references that are necessary to correctly plan the intended action sequence.

Accordingly, among the investigated theories, we will focus on the use of the Frame Semantics

(Fillmore, 1985) and Spatial Semantics (Zlatev, 2007). While the former aims at addressing the problem of scene and event understanding, the latter specifically focuses on the spatial relations involved. It enables a planning and reasoning module to correctly disambiguate objects in the world the robot is acting into. We propose the use of a general structure to represent all the semantics we are interested in. In fact, a typical problem when working with different representations, is that they are totally independent each other. They are not designed to work together in a more general semantic framework. In order to do it, we investigate the use of a new and appealing representation formalism, i.e. *Abstract Meaning Representation* (AMR) (Banarescu et al., 2013). It allows to express semantics without imposing any strong bias to the original sentence or syntactic structure. The final instantiated AMR annotation could be easily mapped to the commands expressed in the robot language (e.g. the logic form), in a way similar to the one proposed in (Thomas and Jenkins, 2012).

In order to prove the effectiveness of the proposed idea, we evaluated existing natural language technologies not customized to the target HRI scenario. In the rest of the paper, in Section 2 different Natural Language Processing tasks are discussed with respect the HRI area. Finally, in Section 3 the conclusion and future works about a new robotic-centric corpus are derived.

2 From Human Voice to Robot Instructions through NLP

A complete NL processing chain for an agent acting in an environment (real or virtual) should be realized as follow: starting from an utterance, a textual representation is obtained from a generic Automatic Speech Recognizer (ASR) (e.g. the CMU Sphinx (Walker et al., 2004)); morpho-syntactic modules (e.g. Stanford CoreNLP (Klein and Manning, 2003)) are then applied; finally, the semantic information is extracted by semantic parsing processors. Starting from this last information, a specific *mapping module* translates the so represented meaning of a sentence in the corresponding robot command.

In this section, different NLP semantic processing tasks useful for robot instructions understanding are described. An evaluation of each task is addressed using 20 commands typical of an HRI scenario. They come from a larger corpus we are labeling to capture major semantic aspects for HRI. These sentences are manually annotated with respect to syntax, part-of-speech tags, parse trees and semantics, with respect to Frame Semantics (Fillmore, 1985) and Spatial Roles (Kordjamshidi et al., 2012). Annotations have been carried out by two of the authors, while conflicts have been resolved by a third one. In the following, a possible NLU pipeline is discussed.

From Voice to Text. The first step in a robot instruction understanding scenario is the automatic transcription of vocal commands. Transcriptions of the utterances are obtained by the audio signal processing performed by ASRs. The ASR engines are usually classified depending on the technique used to generate the Language Model. Two different approaches can be followed for this purpose. The first one, which is called *command-and-control* and is used in the development of several vocal interfaces for commercial systems (i.e. telephone customer care, reservation systems). It requires a grammar-based language specification, typically through Context Free Grammars. The second approach, called *free-form speech*, relies directly on statistical techniques over very large corpora (millions of words), by computing probabilities of sequences of words. While in command-and-control engines it is possible to enrich the grammar with higher-level information, such as attaching semantic information to each rule, in free-form speech engines an external and independent module to compute the desired representation is needed. The use of a grammar-based approach can simplify the semantic parsing process at the expense of coverage, i.e. the constraints imposed to the set of recognized lexicons and utterances. From this point of view, free-form speech systems cover a wider range of linguistic phenomena. For example, in the work of (Thomas and Jenkins, 2012) the official Google speech APIs of the Android environment is used as a free-form speech engine. A *Word Error Rate* of 24% is measured using the Google speech APIs over the 20 test robot commands. It is a promising result, considering that very few sentences are pronounced by English native speakers.

Morphosyntactic Analysis. The last two decades of NLP research have seen the proliferation of tools and resources that reached a significant maturity after 90's. We evaluated a well known platform for the general language processing chain,

that is the Stanford Core NLP platform[1]. It includes tokenization, POS tagging, Named Entity Recognition as well as parsing and is mostly based on statistical, e.g. max-entropy, models for language processing. We want to evaluate the use of these tools to achieve a good command recognition accuracy for a robot. Usually, in NLU morphosyntactic analysis can be crucial to provide features that words alone are not sufficient to express. For example, the dependency parse tree of an utterance could be used in further processing, such as in SRL. We measured the quality of the Stanford parser in terms of *Unlabeled Attachment Score* (UAS) and *Labeled Attachment Score* (LAS) on our 20 test utterances. The former aims at verifying the ability of an algorithm to identify a syntactic relation, while the latter aims at measuring the quality of the relation labeling. We report an accuracy of 87% in UAS and 83% of LAS.

Modeling commands through Semantic Roles. An appropriate theory is necessary in order to capture useful semantics for robot instructions. We argue that Frame Semantics (Fillmore, 1985) could be a good choice to represent different aspects of a robot command. *Frames* are the main structure used to represent and generalize events or actions. They are micro-theories about real world situations (e.g. `movement actions`, such as *moving*, `events`, such as *natural phenomena*, and `properties`, such as *being colored*). Each frame provides its set of semantic roles, i.e. the different elements involved in the situation described by the frame (e.g. an *Agent*). FrameNet (Baker et al., 1998) is a semantic resource reflecting Fillmore's Frame Semantics. In FrameNet lexical entries (such as verbs, nouns or adjectives) are linked to *Frames*, and the roles, expressing the participants in the underlying event, are mapped to *frame elements*. FrameNet has produced an extensive collection of frames as well as a large scale annotated corpus. For example, for the sentence "*take the book on the table*" the following representation is produced: ***take*** [*the book*]$_{Theme}$ [*on the table*]$_{Source}$. In this structure, the different aspects of the TAKING event are highlighted, as the roles THEME and PLACE, suitable for further processing. Frame Semantics can provide a bridge between the linguistic information of a command and its inner robot representation.

We applied Babel, a general purpose SRL system[2] (Croce and Basili, 2011; Croce et al., 2012), to the test sentences. In table 1 results for three different sub-tasks of a SRL chain are reported. In particular, Precision, Recall and F1-Measure are shown for the tasks *Frame Prediction* (FP), *Boundary Detection* (BD), and *Argument Classification* (AC). The first one aims at determining the events evoked in a sentence. The second one is intended to identify the roles involved with respect to a frame. The last one is the task of assigning a label to each role. Performances are lower with respect to the state-of-the-art as, on the one hand, the adopted system was not trained to deal with domain specific phenomena, such as the verb *to be*. On the other hand, the FP badly performed on spoken sentences with jargon expressions, such as "*close the water*", consequently biasing the AC step.

	Precision	Recall	F1-Measure
FP	0.71	0.6	0.65
BD	0.81	0.70	0.75
AC	0.58	0.50	0.54

Table 1: SRL measures on 20 robot commands.

Describing Robot Environment through Spatial Roles. One of the main functions of language is to communicate spatial relationships between objects in the world. Frame Semantics seems inadequate to represent this information at the level of granularity needed by the grounding process of a robotic system. A more specific semantic theory seems thus required and its impact is investigated.

Recently, *Spatial Role Labeling* (SpRL) (Kordjamshidi et al., 2011) was defined as the problem of extracting generic spatial semantics from natural language. The underlying theory is the *holistic spatial semantic theory* (Zlatev, 2007). It defines the basic concepts in the spatial domain of the natural language that help to determine the location or trajectory of motion of a given referent in the discourse. For example, a spatial utterance must address a TRAJECTOR, i.e. the entity whose location is of relevance, or the LANDMARK, i.e. the reference entity in relation to which the location of the trajectory of motion is specified. The SpRL task aims at extracting spatial semantic roles from sentences. Thus, in the sentence "*take the book on the table*", a system should recognize that the preposition "*on*" is the SPATIAL INDICATOR of the relation between "*book*" and "*table*", respectively

[1] `http://nlp.stanford.edu/software/corenlp.shtml`

[2] The system is not domain specific, since it is trained on the FrameNet 1.5 dataset

	Precision	Recall	F1-Measure
SI	0.78	0.84	0.81
TR	0.80	0.61	0.70
TD	0.75	0.75	0.75

Table 2: Spatial Role Labeling results.

a TRAJECTOR and a LANDMARK. These information should help a robotic system to correctly determine which book has to be taken within the physical world, i.e. the one on the table. In table 2 we report performance measures in terms of Precision, Recall and F1-Measure of a Spatial Role Labeler (Bastianelli et al., 2013).These results refer to the SPATIAL INDICATOR (SI), TRAJECTOR (TR) and LANDMARK (LD) (Kordjamshidi et al., 2011) labeling on the 20 test sentences used above.

Expressing Rich Semantic Information through AMR. In order to integrate the information conveyed by the Frame Semantics and the Spatial Semantics, we want to propose a representation flexible and as much as possible close to the domotic domain, i.e. the robot language. The Abstract Meaning Representation (AMR) (Banarescu et al., 2013) is a novel semantic representation language that allows to represent semantics in an abstract way, focusing on concepts, their instances and the relations among them. According to this notation, the meaning of a sentence is represented as a rooted, labeled (acyclic) graph, where the semantic structure is built in a recursive way. In AMR, sentences that have different syntactic structures but basically the same meaning are represented by the same structure. While the AMR proposed in (Banarescu et al., 2013) uses the PropBank frame sets (Palmer et al., 2010), we want to adopt it to embed the semantics coming both from Frame Semantics and Spatial Semantics. The command "*take the book on the table*" will be represented as follows:

```
(t / take − Taking
    :Theme(b / book)
    :Source(t1 / table)
    :location(o / on
            :trajector(b))
            :landmark(t1)))
```

Here, *book* and *table* represent concepts; `b` and `t1` are the instances respectively related. Frame Semantics is represented by the instance `t` of the verb *take*, evoking the frame `Taking`. In a similar way, the two semantic roles `Theme` and `Source` are defined as the instances `b` and `t1`. The spatial relation `location` is defined across the two semantic roles, linking the `b` instance to the `t1` through the preposition *on*. This structure appears to be very agile for computing and for the HRI interface design. It can be seen as the abstraction step in the representation of meaning, used before the final translation into the logic-like formalism. This latter is closer to the robot representation, but more complex to manage. The tree-like structure of AMR makes it very easy to navigate, elaborate and visualize. Furthermore, many consolidated formalism can be derived from this one, as neodavidsonian Discourse Representation Structures (Bos and Oka, 2007). While DRSs are closest to a possible representation of the world a robot might have, AMR offers a promising degree of abstraction, especially because we want to follow a data-driven approach, without relying on too rigid representations or tools. It seems to embed in a logic-like formalism all the information needed for the symbol grounding process of a robot, such as relation between linguistic objects as well as roles. Actually, a mapping procedure to compile the final AMR representation is under development.

3 Conclusion and Future Work

In this paper, we discussed the possibility of combining state-of-the-art textual inference technologies in the design of HRI architectures. Moreover, we experimented standard NL inference tools to verify the quality achievable by current technologies. This is the first step of a research that aims at defining a unified framework able to capture the major semantic aspects of linguistic utterances within the HRI field. Clearly, many aspects of this challenging research area are underway. A deeper investigation of the semantic theories and representation schemata is still needed. As we are interested in data driven paradigms, we need to improve the adaptation capability of existing technologies and to provide more labeled data for them. At the moment, we collected about 450 audio streams (recorded during the Robocup 2013) expressing generic robot commands from different speakers. We are starting labeling them according to the semantic theories investigated in this paper. We are planning to release the annotated resource, as soon as a significant amount of annotated sentences has been produced. Further evaluations are finally needed to investigate the impact of the error rate through the entire pipeline.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, Association for Computational Linguistics '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August. Association for Computational Linguistics.

Emanuele Bastianelli, Danilo Croce, Daniele Nardi, and Roberto Basili. 2013. Unitor-hmm-tk: Structured kernel-based learning for spatial role labeling. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval-2013)*, Atlanta, Georgia, USA, June.

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

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

Danilo Croce and Roberto Basili. 2011. Structured learning for semantic role labeling. In *AI*IA*, pages 238–249.

Danilo Croce, Giuseppe Castellucci, and Emanuele Bastianelli. 2012. Structured learning for semantic role labeling. *Intelligenza Artificiale*, 6(2):163–176.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine*, 31(3).

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

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of ACL'03*, pages 423–430.

Parisa Kordjamshidi, Martijn Van Otterlo, and Marie-Francine Moens. 2011. Spatial role labeling: Towards extraction of spatial relations from natural language. *ACM Trans. Speech Lang. Process.*, 8(3):4:1–4:36, December.

Parisa Kordjamshidi, Steven Bethard, and Marie-Francine Moens. 2012. SemEval-2012 task 3: Spatial role labeling. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the Main Conference and the Shared Task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012),*, pages 365–373. Association for Computational Linguistics, June.

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

Beth Levin. 1993. *English Verb Classes and Alternations A Preliminary Investigation*. University of Chicago Press, Chicago and London.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. *Semantic Role Labeling*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Brian J Thomas and Odest Chadwicke Jenkins. 2012. Roboframenet: Verb-centric semantics for actions in robot middleware. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 4750–4755. IEEE.

Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, and Joe Woelfel. 2004. Sphinx-4: A flexible open source framework for speech recognition.

Jordan Zlatev. 2007. Spatial semantics. *Handbook of Cognitive Linguistics*, pages 318–350.