Unsupervised Online Grounding for Social Robots (Extended Abstract)[*]

Oliver Roesler[1] and Elahe Bagheri[2]

[1] Artificial Intelligence Lab, Vrije Universiteit Brussel, Brussels, Belgium
oliver@roesler.co.uk

[2] Robotics and Multibody Mechanics Research Group, Vrije Universiteit Brussel and Flanders Make, Brussels, Belgium elahe.bagheri@vub.be

1 Introduction

Robots that incorporate social norms in their behaviors are seen as more supportive and friendly. Since it is impossible to manually specify the most appropriate behavior for all possible situations, robots need to be able to learn it through trial and error, by observing interactions between humans, or by utilizing theoretical knowledge available in natural language. In contrast to the former two approaches, the latter has not received much attention because understanding natural language is non-trivial and requires proper grounding mechanisms to link words to corresponding perceptual information. Previous grounding studies have mostly focused on grounding of concepts relevant to object manipulation [1,4], while grounding of more abstract concepts relevant to the learning of social norms has so far not been investigated.

In this paper, we present an unsupervised cross-situational learning based online grounding framework to ground emotion types, emotion intensities and genders through their corresponding concrete representations, which represent sets of invariant perceptual features obtained through an agent's sensors that are sufficient to distinguish percepts belonging to different concepts [3], extracted from audio with the help of deep learning. The proposed framework is evaluated through a simulated human-agent interaction experiment in which the agent listens to the speech of different people and receives at the same time a natural language description, describing the gender of the observed person as well as the experienced emotion. Furthermore, the proposed framework is compared to a Bayesian grounding framework that has been employed in several previous studies to ground words through a variety of different percepts [1,4].

2 System Overview

The employed grounding framework consists of three parts: (1) Perceptual feature extraction component, which extracts audio features from video using openEAR [2], (2) Perceptual feature classification component, which uses deep neural networks to obtain concrete representations of perceptual features, (3) Language grounding component, which identifies auxiliary words, i.e. words that have no corresponding concrete representations, and creates mappings from non-auxiliary words to corresponding concrete representations using cross-situational learning.

[*] This is an extended abstract of Roesler and Bagheri [5].

3 Results

The obtained results show that the framework is able to identify auxiliary words and ground non-auxiliary words, including synonyms, referring to abstract concepts through their corresponding emotion types, emotion intensities and genders. Furthermore, they illustrate that the grounding algorithm employed by the proposed framework depends on the accuracy of the used concrete representations, which are in this study obtained through deep learning, but does not require perfectly accurate representations because the framework is already able to obtain all correct mappings, if the accuracy of the concrete representations is on average only around 85% for all considered modalities. Additionally, the proposed framework outperformed the baseline framework in terms of the accuracy of the obtained groundings as well as its ability to learn new groundings and continuously update existing groundings during interactions with other agents and the environment, which is essential when considering real-world deployment. Finally, the framework is also more transparent, due to the creation of explicit mappings from words to concrete representations.

4 Conclusion

The proposed framework allowed identification of auxiliary words and grounding of abstract concepts, like emotion types, emotion intensities and genders, through their corresponding concrete representations in an online manner using cross-situational learning. In future work, we will integrate the framework with a knowledge representation to explore the utilization of abstract knowledge to increase the sample-efficiency of the grounding mechanism as well as the accuracy of the obtained groundings, and enable agents to reason about the world with the help of an abstract but grounded world model.

References

1. Aly, A., Taniguchi, T.: Towards Understanding Object-Directed Actions: A Generative Model for Grounding Syntactic Categories of Speech through Visual Perception. In: IEEE International Conference on Robotics and Automation (ICRA). Brisbane, Australia (May 2018)
2. Eyben, F., Wöllmer, M., Schuller, B.: OpenEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. In: Proceedings of the 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops. Amsterdam, Netherlands (September 2009)
3. Harnad, S.: The Symbol Grounding Problem. Physica D **42**, 335–346 (1990)
4. Roesler, O.: Unsupervised Online Grounding of Natural Language during Human-Robot Interaction. In: Second Grand Challenge and Workshop on Multimodal Language at ACL 2020. Seattle, USA (July 2020)
5. Roesler, O., Bagheri, E.: Unsupervised Online Grounding for Social Robots. Robotics **10**(2) (April 2021). https://doi.org/https://doi.org/10.3390/robotics10020066