Enhancing Unsupervised Natural Language Grounding through Explicit Teaching

Oliver Roesler
Artificial Intelligence Lab
Vrije Universiteit Brussel
Brussels, Belgium
oliver@roesler.co.uk

https://doi.org/10.31256/Bf9Vw8C

***Abstract*—In this paper, a grounding framework is proposed that combines unsupervised and supervised grounding by extending an unsupervised grounding model with a mechanism to learn from explicit human teaching. To investigate whether explicit teaching improves the sample efficiency of the original model, both models are evaluated through an interaction experiment between a human tutor and a robot in which synonymous shape, color, and action words are grounded through geometric object characteristics, color histograms, and kinematic joint features. The results show that explicit teaching improves the sample efficiency of the unsupervised baseline model.**

***Index Terms*—language grounding, cross-situational learning, sample efficiency, human-robot interaction**

I. Introduction

The need for robots that are able to understand natural language instructions is growing due to an increasing number of service robots that are employed in human-centered environments. To this end, connections between words and percepts need to be created through grounding because language only has meaning, if it is linked to the physical world [1].

Previous studies that investigated grounding employed either unsupervised [2]–[4] or supervised [5], [6] approaches. The former have the advantage that no human tutor is required for grounding, however, they require a large number of situations to learn the correct grounding, i.e. they are less sample efficient, and are often also less accurate than supervised approaches. In comparison, the latter are often more accurate and can already learn the correct mappings from a very small number of situations, however, they do not work in the absence of a human tutor.

In this paper, both approaches are combined by extending a recently proposed unsupervised cross-situational learning based grounding framework [7], [8] to learn from explicit human teaching, if available. The main aim is to investigate whether this extension increases the model's sample efficiency, i.e. whether it reduces the number of interactions required until the model obtains the correct mappings between words and percepts.

The rest of this paper is structured as follows: Section (II) describes the extended grounding framework. The experimental design and obtained results are described in Sections (III and IV). Finally, Section (V) concludes the paper.

II. System Overview

The used grounding system consists of the following parts:

1) **3D object segmentation system**, which employs a model based 3D point cloud segmentation approach [9] to segment objects into point clouds. The shapes and colors of objects are represented through Viewpoint Feature Histogram [10] descriptors, which represent the object geometries taking into consideration the viewpoints, while ignoring scale variances, and color histograms.
2) **Action recording system**, which records the vertical position of the robot's torso, the angles of the arm flex and wrist roll joints, the velocity of the robot's base and the binary state of the gripper, i.e. open or closed, during action execution. The recorded data is then combined into an action feature vector, which represents the change of the recorded characteristics between the beginning and the end of an action.
3) **Percept clustering component**, which converts percepts to abstract representations through clustering to enable the CSL algorithm to use them to ground encountered words as proposed by [11]. The used clustering algorithm is DBSCAN [12] because it does not require the number of clusters to be specified in advance, which is important since it cannot be assumed that the number of percepts is known in advance. Cluster numbers were calculated prior to grounding so that they could be provided to the CSL algorithm. Shape, color, and action percepts achieved mean adjusted rand scores [13] of 0.84, 1.0, and 0.99, respectively.
4) **Language grounding component**, which uses an extended version of the cross-situational learning based grounding algorithm proposed by [8]. The original grounding algorithm grounds words and phrases through cluster numbers of percepts in an unsupervised manner without being able to take into account any teaching or feedback by a human tutor. Thus, in this study an extension is proposed that provides a mechanism to learn mappings from explicit teaching. The new mechanism uses similar to the original one cross-situational learning to determine the correct mappings, however, it requires the tutor to artificially create a situation where only one percept occurs twice and only one word is given

to the robot, which should be grounded through that percept (Section III). When a new mapping has been obtained through explicit teaching, it is added to the set of previously obtained mappings, which also includes mappings obtained through the unsupervised grounding algorithm during regular situations.

III. Experimental Setup

During the experiment a human tutor and HSR robot [14] interact in front of a table with one or two objects on top of it. Interactions can be of two types. Either the human tutor asks the robot to perform an action on the object or the tutor tries to teach the robot the correct mapping for a shape, color, or action word. The former interactions use the following procedure.

1) The tutor places an object on the table and the robot determines the corresponding shape and color percepts.
2) The human tutor provides an instruction to the robot.
3) The human tutor teleoperates the robot to execute the action provided through the instruction and the robot determines a corresponding action percept.
4) The robot employs clustering to convert all encountered percepts to abstract representations.
5) Words are grounded through obtained cluster numbers by the CSL based grounding algorithm.

In contrast, situations in which the human tutor tries to teach the robot a specific mapping follow the following procedure.

1) The human tutor places two objects, which have either the same shape or color, on the table and the robot determines the corresponding percepts, if the tutor tries to teach the correct mapping for a shape or color word. Otherwise, to teach an action word, the human tutor places two objects on the table that have different shapes and colors and executes the same action for both of them so that only the action percept occurs twice.
2) The human tutor provides one single word, which refers to the percept that occurs twice.
3) The robot creates a corresponding mapping and adds it to the set of previously obtained mappings.

To create the 2,500 situations used in this study without having to perform 2,500 interactions, the following procedure is applied. First, a total of 125 interactions are performed to record perceptual information for all combinations of employed shapes, colors, and actions, while skipping the last two steps of the interaction procedure, i.e. steps 4 and 5. Afterwards, all possible unique sentences are obtained by creating all possible combinations of shape, color, and action words. Finally, each sentence is randomly assigned one shape, color, and action percept that correspond to the words in the sentences, leading to overall 2,500 situations.

Each sentence has the following structure: "*action* the *color shape*", where *action*, *color*, and *shape* are replaced by one of their corresponding words. Each action and color can be referred to by two different words, e.g. the color green can be referred to by "green" or "greenish", while each shape has five corresponding words, e.g. "latte", "milk", "milk tea", "coffee" or "espresso" for cup.

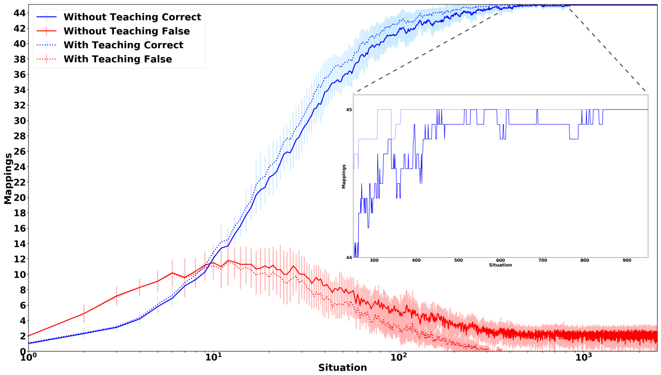

Fig. 1. Grounding results showing means and standard deviations of correct and false mappings over all 2,500 situations encountered by the robot for 10 different sequences. The dotted lines represent the results, when the tutor teaches the robot a correct mapping after on average every 9 situations, while the continuous lines represent the results, when no explicit teaching is provided.

IV. Results and Discussion

The results show that teaching increases the convergence towards the correct mappings (Figure 1). If no teaching is provided, the algorithm requires about 850 situations to ground all words correctly for all 10 different sequences, while it only requires about 350 situations, when the human tutor explicitly teaches one mapping to the robot after on average every 9 situations. Before obtaining all correct mappings teaching also leads to a slightly higher number of correctly grounded words. Even after all words are correctly grounded, the number of false mappings oscillates around 2, when no teaching is provided, because the algorithm allows words to be grounded through several percepts to handle homonyms. In contrast, no false mappings are obtained in case of the extended model. If the human tutor teaches all 45 words at the beginning of the experiment, the model learns all words after 45 situations, assuming that all encountered percepts are correctly clustered. While teaching all words explicitly is possible for the small number of words used in this scenario, it would not be feasible for a much larger number of words, which illustrates the importance of the unsupervised grounding mechanism.

V. Conclusions and Future Work

An unsupervised grounding model was extended to allow it to benefit from explicit teaching by a human tutor. The proposed model was evaluated through a human-robot interaction experiment and compared to the original model that does not allow explicit teaching. The results showed that with teaching the model grounds all words on average about 2.5 times faster than without teaching. In future work, the model will be evaluated for longer and more complex sentences that contain a larger number of words. Furthermore, the influence of wrong teaching, i.e. when the tutor on purpose or by accident provides a wrong word during teaching, will be investigated. Finally, the model will be extended to allow human feedback for already obtained groundings.

References

[1] S. Harnad, "The symbol grounding problem," *Physica D*, vol. 42, pp. 335–346, 1990.

[2] C. R. Dawson, J. Wright, A. Rebguns, M. V. Escárcega, D. Fried, and P. R. Cohen, "A generative probabilistic framework for learning spatial language," in *IEEE Third Joint International Conference on Development and Learning and Epigenetic Robotics (ICDL)*, Osaka, Japan, August 2013.

[3] O. Roesler, A. Aly, T. Taniguchi, and Y. Hayashi, "A probabilistic framework for comparing syntactic and semantic grounding of synonyms through cross-situational learning," in *ICRA-18 Workshop on Representing a Complex World: Perception, Inference, and Learning for Joint Semantic, Geometric, and Physical Understanding.*, Brisbane, Australia, May 2018.

[4] ——, "Evaluation of word representations in grounding natural language instructions through computational human-robot interaction," in *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Daegu, South Korea, March 2019.

[5] L. Steels and M. Loetzsch, "The grounded naming game," in *Experiments in Cultural Language Evolution*, L. Steels, Ed. Amsterdam: John Benjamins, 2012, pp. 41–59.

[6] L. She, S. Yang, Y. Cheng, Y. Jia, J. Y. Chai, and N. Xi, "Back to the blocks world: Learning new actions through situated human-robot dialogue," in *Proceedings of the SIGDIAL 2014 Conference*, Philadelphia, U.S.A., June 2014, pp. 89–97.

[7] O. Roesler and A. Nowé, "Simultaneous action learning and grounding through reinforcement and cross-situational learning," in *ALA 2018, Adaptive Learning Agents Workshop.*, Stockholm, Sweden, July 2018.

[8] ——, "Action learning and grounding in simulated human robot interactions," *The Knowledge Engineering Review*, vol. 34, no. E13, November 2019.

[9] C. Craye, D. Filliat, and J.-F. Goudou, "Environment exploration for object-based visual saliency learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, Stockholm, Sweden, May 2016.

[10] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu, "Fast 3D recognition and pose using the viewpoint feature histogram," in *Proceedings of the 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Taipei, Taiwan, October 2010, pp. 2155–2162.

[11] O. Roesler, "A cross-situational learning based framework for grounding of synonyms in human-robot interactions," in *In Proc. of the Fourth Iberian Robotics Conference (ROBOT)*, Porto, Portugal, November 2019.

[12] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, Portland, Oregon, USA, August 1996, pp. 226–231.

[13] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, December 1985.

[14] *HSR Manual*, 2017th ed., Toyota Motor Corporation, April 2017.