Enhancing Unsupervised Language Grounding through Online Learning

Oliver Roesler[1]

***Abstract*— To enable natural human-robot collaboration robots need to understand natural language, which in turn requires mechanisms to ground language by connecting words to corresponding percepts. This paper, presents an unsupervised online grounding framework for grounding of synonymous words through corresponding percepts. The framework is evaluated through a human-robot interaction experiment and compared to a Bayesian grounding framework, which requires an offline training phase. The results show that the proposed framework outperforms the baseline, thereby illustrating the benefit of online learning for natural language grounding.**

I. Introduction

The number of robots that work closely with humans is growing, thereby, increasing the demand for robots with proper communication abilities. Since language is the most natural form of communication for humans, robots need to understand the meaning of language, which requires words to be linked to the physical world [1]. Many different approaches have been proposed to create these links and thereby ground words through corresponding percepts. In general, they can be separated into supervised approaches, e.g. [2], that require the presence of a teacher to provide demonstrations or feedback and unsupervised approaches, e.g. [3], that do not require any explicit teaching or feedback, but utilize co-occurrence information to identify through which percepts a word is grounded. The majority of previous studies that investigated unsupervised grounding employed algorithms that only work offline, i.e. the algorithms need to be trained before deployment, which makes these algorithms unsuitable for real-time human-robot interactions.

In this paper, a cross-situational learning based unsupervised grounding framework is presented that does not require an explicit training phase, but instead updates the obtained co-occurrence information continuously with every new situation. For every situations the framework converts obtained percepts to abstract representations through clustering and provides them afterwards to a cross-situational learning algorithm to identify the corresponding percepts for each word. The grounding performance of the presented framework is evaluated through a human-robot interaction experiment and compared to the grounding performance of the Bayesian grounding framework presented in [4].

The remainder of this paper is organized as follows: Sections (II and III) describe the framework and experimental design. Section (IV) presents the obtained results and Section (V) concludes the paper.

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

II. Grounding Framework

The grounding framework consists of below four parts:

1) **3D object segmentation system**, which obtains the shapes and colors of objects through a model based 3D point cloud segmentation approach. Colors are represented through histograms and shapes are represented through Viewpoint Feature Histogram [5] descriptors, which represent the object geometries taking into account the viewpoints, while ignoring scale variances.
2) **Action recording system**, which creates action feature vectors representing the change of 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, between the beginning and end of an action.
3) **Percept clustering component**, which enables the cross-situational learning based algorithm to ground words through an abstract representation of percepts obtained through clustering as proposed by [6]. Since the number of clusters might not be known in advance, DBSCAN [7] is used for clustering because it does not require the number of clusters to be specified in advance. Clusters are re-calculated for every situation to ensure that new percepts are incorporated.
4) **Language grounding component**, which uses a modified version of the cross-situational learning based grounding algorithm proposed in [8] to connect words and corresponding percepts in an unsupervised manner as well as detect auxiliary words, which are words that have no corresponding percept and only exist for grammatical reasons.

III. Experimental Setup

In the employed experimental scenario, which is based on the scenario used in [4], a human tutor and robot interact in front of a table according to below procedure.

1) The human tutor places an object on the table and the robot determines the object's geometric characteristics and color to create corresponding feature vectors.
2) An instruction, which describes how to manipulate the object, is given to the robot by the human tutor.
3) The human tutor teleoperates the robot to execute the action provided through the instruction, while several kinematic characteristics are recorded and converted into an action feature vector.

A total of 125 interactions were performed to record perceptual information for all combinations of the 5 employed shapes, colors, and actions. Since instruction words were

selected randomly for each situation, except that words had to fit the encountered percepts, their number of occurrences in the data varies between one and fourteen times. Each sentence has the following structure: "(please) *action* the *color shape*", where *action*, *color*, and *shape* are substituted by one of their corresponding words, while the word "please" is optional and occurs for less than half of the sentences. Each action and color are referred to by two different synonymous words and each shape by five synonyms, e.g. bottle is referred to by "coca cola", "soda", "pepsi", "coke" and "lemonade". Ten different interaction sequences were created, by randomly changing the order of the recorded situations, to ensure that the grounding performance is not due to the specific order in which situations are encountered. During training and testing the obtained situations are given to the proposed and baseline framework. The former gets training situations separately one after the other to simulate real-time processing, i.e. it only has access to the percepts and words of the current and previous situations, while the baseline framework receives all training situations, i.e. corresponding sentences and percepts, at once, since it is not able to learn online.

IV. Results and Discussion

Although the presented framework does not require an explicit training phase, all ten different sequences of situations were split into a training (first 60%) and a test set (last 40%) because the baseline framework does not work online and requires an explicit training phase. Figure (1) shows that the presented model achieved accuracies of more than 95% for all modalities and nearly 80% for auxiliary words. The standard deviations illustrate that the grounding performance is influenced by the composition of the training and test sets. Interestingly, when all situations are used for training the presented framework achieves correct grounding for all words, which would be the normal case, since it does not require explicit training and continues learning after deployment. In comparison, the baseline model achieves mean accuracies of about 55% and 50% for action and color words, while it only achieves accuracies of 25% and less than 5% for auxiliary and shape words, respectively. The low accuracy for shape words might be due to their relatively low number of occurrences due to the larger set of synonyms, since each shape has five synonyms, while each color and action has only 2 synonyms. This is also supported by the fact that the grounding accuracy for shape words significantly improves, when all situations are used for training and testing.

Overall the results show that the presented unsupervised grounding framework achieves better grounding performance than the baseline model, which has been used in previous studies for similar grounding scenarios, even though the former was artifically limited to only learn from the training situations to ensure a fair comparison. Therefore, the results also illustrate the benefit of online learning for language grounding because it prevents that models are limited by the situations encountered during training.

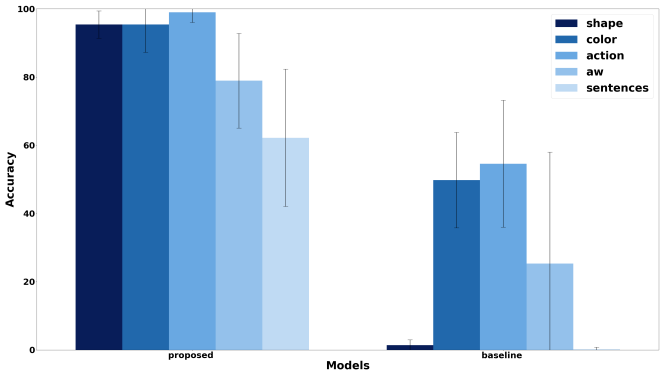

Fig. 1: Mean grounding accuracy results and corresponding standard deviations for both models, all modalities and complete sentences.

V. Conclusions and Future Work

An unsupervised cross-situational learning based online grounding framework was presented, evaluated through a human-robot interaction experiment and compared to a Bayesian grounding model that has previously been used for similar grounding scenarios and requires an offline training phase. The results show that despite being restricted to only learn from a subset of the situations the presented framework outperforms the baseline, thereby illustrating the benefit of online learning for language grounding. In future work, the framework will be evaluated for more complex sentences that contain a larger number of words as well as homonyms. Furthermore, 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] 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.

[3] 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*, vol. 32, no. 4, p. 6476, 2011.

[4] O. Roesler, A. Aly, T. Taniguchi, and Y. Hayashi, "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] 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.

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

[7] 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.

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