GLU 2017 International Workshop on Grounding Language Understanding
25 August 2017, Stockholm, Sweden

Grounding Imperatives to Actions is Not Enough: A Challenge for Grounded NLU for Robots from Human-Human Data

Julian Hough, Sina Zarrieß, David Schlangen

Dialogue Systems Group // CITEC
Faculty of Linguistics and Literature
Bielefeld Univerisity
firstname.lastname@uni-bielefeld.de

Abstract

We present a proposal for a Natural Language Understanding method for simple pick-and-place robots which maps utterances to different levels in an action hierarchy. The hierarchy is a graph containing both lower-level action and higher-level goal levels. This attempts to overcome the surprising lack of overt imperative verb forms in natural task-oriented dialogue, which we show to be the case statistically in a human-human corpus. This proposal shifts the task away from mapping utterances to either actions or goals exclusively, and instead allows flexible mapping to both actions and goals during the interaction. We also show how a continuous communicative grounding mechanism is vital for achieving fluid interaction and show how confirmations and repairs can refer to both the goal and action levels, and that reliance on these overt signals of understanding alone is inadequate for a natural model.

Index Terms: Grounded NLU, actions, verb phrase ellipsis

1. Introduction

Natural Language Understanding (NLU) for robots cannot currently facilitate the natural level of interaction found in human-human data. Taking a simple pick-and-place domain of building shapes from Pentomino pieces as an example, we would eventually want a system to have the level of understanding ability which B shows when instructed by A in (1).

(1)

1.	A:	We're going to build a pyramid.
	B:	Okay
	A:	First, we're going to make the bottom corner.
2.	A:	Take the green W.
	B:	[takes green W]
	A:	Right.
	A:	Now rotate it 90 degrees to the right.
	B:	[turns green W 90 degrees to right]
	A:	No, sorry, to your left.
	B:	[turns green W 160 degrees to left]
	A:	A bit more.
	B:	[turns green W 20 degrees to left]
	A:	And put it in the bottom left corner.
	B:	[puts it in bottom corner]
3.	A :	The green W goes in the bottom left corner facing up.
	B :	[picks up green W] [rotates green W 90 degrees]
	B :	[places green W in bottom left corner]
4.	A:	Now take the green T

This interaction highlights the present challenges. In the first section in (1), a higher-level *goal* is communicated by A followed by an initial attempt to communicate the first *sub-goal* of the task. In the second section, a series of lower-level instructions on the action level are given, including *confirmation* of a successfully completed action and *repair* of an initially miscommunicated action.

In the repair "No, sorry, to your left", the action is not re-referenced explicitly but a bare fragment is used. The verb can be omitted here because of the mutually available context available to the interaction participants that the fragment refers to an action in progress. We will call this phenomenon *embodied verb phrase ellipsis*, and as we will show in §5 that this is in fact extremely frequent in such a domain.

An alternative to the second section is in the third section where a more complete description of the goal state is given rather than action imperative forms. In the last section, without confirmation of B's success, the instructor A continues with the next sub-goal, showing that to achieve communicative grounding in the sense of [1], one cannot rely on the presence of conventional confirmation signals.

In addition to these interpretation challenges, the instruction follower (IF)'s actions will often overlap with the instruction giver (IG)'s speech, and, when appropriate, the IF will take initiative and not in fact be a follower but a leader.

While several methods have been developed for learning the grounding of instructions into logical forms for a robot to carry out a plan [2, 3], these do not allow the flexibility required for the type of interaction in (1) and rely on explicit verb forms which are directly grounded in a corresponding action. Even if statistical NLU methods allow for some flexibility in the form, these still only permit a command-and-control Human-Robot Interaction (HRI) with long waiting times and no ability to adjust plans on the fly. In this paper, we outline a model which begins to address this restrictive reliance on overt verb forms and sentential commands in §2-4 and then we show the extent of the remaining challenges briefly in a small corpus study of human-human interactions in §5.

2. Towards fluid, interactive NLU for pick-and-place HRI

For the long term goals of this current work, we want to move towards grounded NLU for HRI with the following properties:

- Implicit reference to action can be resolved without the need for an overt verb form.
- Goal-referencing as well as action-referencing utterances can be interpreted to make decisions about the next action.
- Interpretation that the robot has achieved the desired goals (communicative grounding) can be implicit, using context, without needing to rely on explicit confirmation utterances like 'yes' or 'that's correct'.
- Repairing or modifying the robot's current action can be interpreted online and in as fluid a manner as possible.

10.21437/GLU.2017-18

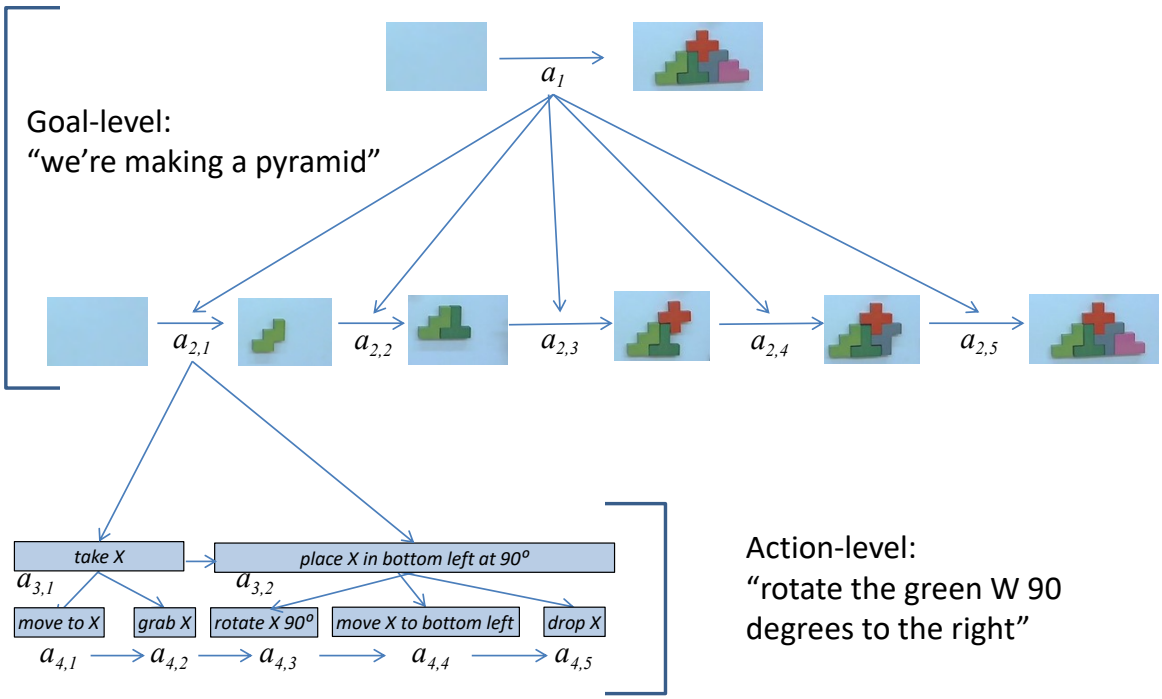

Figure 1: *A multi-layered hierarchical representation of a puzzle construction task. Vertical arrows indicate grounded in links whereby the higher level actions are grounded in the lower-level ones. Horizontal ones mean time-linear dependency. This is one possible route to solving the puzzle which the robot may take. Instructions may be grounded in different layers, from higher-level to lower-level.*

To address the above, we provide two formal proposals here. Firstly, we show that when intentions are characterized as parts of a hierarchical plan, the utterances in (1) can become interpretable. Secondly, we present a communicative grounding model, developed from [4], which shows how grounding can happen fluidly in HRI. We finish by showing how the model applies to data from human-human interaction and discuss the outlook.

3. HRI Intentions as Adjustable Hierarchical Action Graphs

Firstly, we follow [3] in showing how a hierarchical structure can capture simple robotic tasks in a useful way for NLU. Fig. 1 shows the decomposition of the task of building a pyramid from Pentomino blocks as per (1).

Here we decompose the task into different action levels. The top-level action in the graph a_1 is the overall goal of building the pyramid, as per the first section of interaction (1). We will call the second-level actions the *sub-goal* level, constituting tasks which must be completed in a given time-linear order to achieve the goal. For the purposes of the discussion here, we group the two lowest levels into one *action* level, as they represent the low-level manipulation actions which do not require representation of the higher-level goals. In terms of communicating this complex task through language, it can be said that the higher-levels can be *grounded in* the lower-levels [5].

From the interpreter/robot's perspective, the task is to find the best match for the incoming words to a node in this graph. If the robot is able to compute on the goal and sub-goal levels, it should be able to plan the lower-level actions to achieve them, taking the initiative based on its knowledge of the task. If it fails to interpret the correct goals, repair on the lower action-level can also be interpreted.

We propose a mixed and flexible strategy of interpretation. A robot which can only ground instructions to the lower-level actions will make for a tedious and inflexible interaction. On the other hand, a robot that can only understand high-level plans lacks flexibility in terms of online adjustment of the sub-goals that might be required in real-world, dynamic situations.

4. Continuously Grounding Intentions

The task of computing the user's current intention from the user's speech and the current state as a node on the action graph becomes more complex during mis-communication. To be able to recover from computing an inappropriate intention, there must be a capacity to recognize *repairs*, and also, when appropriate, the evidence that things are (back) on track, either through interpreting a confirmation or a tacit sign of the user committing to the robot's action.

For this we require a continuous communicative grounding model, in the sense of [1]. Due to space constraints we direct the reader to [4], however the essence of the model is that it consists of two parallel state charts, one for the robot (observed) and one for the user (estimated), where for the most informative current intention that can be recognized by the robot, each agent goes through different stages of commitment to it from 'uncommitted' to 'showing_commitment', to 'committed'. There is also a 'repairing' state for both agents. The 4 states in action can be seen in Fig. 2 in the course of a repair interaction with a simple pick-and-place robot. Both states $Robot$ and $User$ end up being 'committed' (and consequently grounded) after a repair interaction.

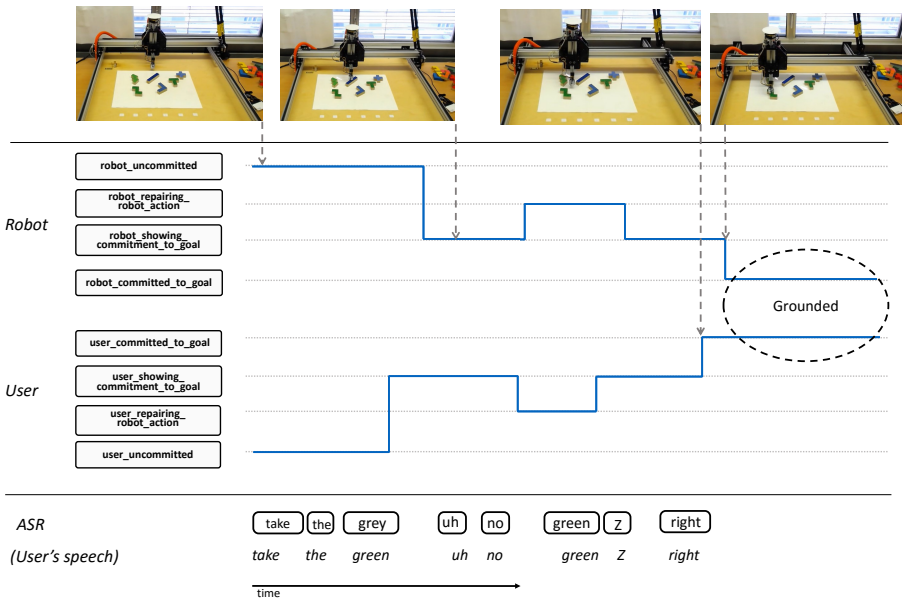

Figure 2: *Concurrent User and Robot grounding states during an interaction where an initial mis-recognition of 'green' as 'grey' by the ASR, and confusion over colours in reference resolution where 'grey' gives higher probability to a blue object. The recognition of repair allows the participants to become grounded again.*

5. Corpus Study on Grounded Verb Usage and Grounding Acts

Given the two proposals of characterizing intentions in HRI as hierarchical action graphs and a continuous grounding state machine, we would like to see how these play out in reality in human-human interactions.

We attempt to answer the following questions from human-human data:

- How often are imperative verbs detailing the lowest level actions used, verses goal-level descriptive verbs?
- How often are bare forms used (no verb at all), and at what points in the interaction?
- What is the distribution of grounding acts (confirmations, repairs) in the corpus and how often is this done implicitly?

5.1. Methodology

We use the PENTO-CV corpus from the `PentoRef` release [6],[1] a German corpus of situated interactions wherein 8 pairs of 2 participants instruct one another via video and audio feed to manually complete a Pentomino puzzle. Both participants have a turn at playing one of two roles: the *instruction giver* (IG) is given a photograph of the final goal configuration of the puzzle pieces and can see the puzzle being constructed by the *instruction follower* (IF). Audio access is full-duplex and bidirectional while only the IG can see the hand actions of the IF as they construct the puzzle.

Every utterance from each participant was segmented and transcribed and each hand action was annotated according to one of the following action labels with higher-level actions in brackets, in addition to the identifier of the piece being manipulated:

```
move_to_piece (take) (put)
grab_piece (take) (put)
move_with_piece (place) (put)
rotate_piece (place) (put)
drop_piece (place) (put)
retract
```

Using one of the dialogues to build and test a simple automatic dialogue act tagger which used the action labels of the hand actions and the words as features, we tagged each utterance with one of the dialogue acts in Fig. 3.

We tagged all the verbs in the corpus manually in terms of their type: *action-level* or *goal-level*, whereby the former consists of direct imperatives to manipulation instructions (e.g. '*take* the red cross') while the goal-level verbs describe the final desired state (e.g. 'the red cross *sits* in the corner'). An example of the annotation scheme applied to a section of one of the interactions is in Fig. 4– this is just the instruction giver's speech, as the follower did not make a verbal contribution in this section.

5.2. Results

Our preliminary results for the verb distribution over *instruct* and *extend* acts can be seen in Table 1. As can be seen, in extend acts, the number of overt verb forms used is much lower, with the number of action-level forms being marginally more prevalent than goal-level forms. With only just over 40% containing overt verb forms, extensions are often bare prepositional phrases or modifiers and provide a substantial challenge in the resolution of the appropriate action. We term this phenomena *embodied verb phrase ellipsis*, and we hope it will be taken seriously by system designers.

As for the distribution of the grounding acts, as can be seen in Table 2, on average slightly less than one dialogue act trig-

[1] Our dataset is available from `https://github.com/dsg-bielefeld/pentoref`, here using release `1.0`.

Dialogue Act	Description
`instruct`	The utterance up to, and overlapping with, the beginning of the first attempt at placing a piece.
`extend`	Any subsequent utterance which continues beyond the first attempt at the subgoal without overt repair.
`repair`	An overt repair signal with the majority of words negative discourse markers like 'nee', 'nein' etc.
`confirm`	An overt confirmation with confirmatory words such as 'ja', 'korrekt' etc.

Figure 3: *Dialogue act mark-up*

Instruction Giver Utterance	English translation	Verb level	Dialogue act
dann kommt das pinke	then comes the pink	goal-level	instruct
das ist der Unterkiefer	this is the mandible	goal-level	instruct
der kommt dann in die Ecke rein	this then goes in the corner	goal-level	instruct
einmal nach links	once to the left	None	extend
nee	no	None	repair
liegenlassen	put it down	action-level	extend
nach links schieben und oben einpassen	move it to the left and fit it in the top	action-level	extend

Figure 4: *Example dialogue stretch with annotations.*

% utts containing	instruct	extend
action-level verb	33.1	23.7
goal-level verb	33.2	21.2
any verb	61.7	42.5

Table 1: *Occurrences of different verb types in the Instruction Giver's speech*

Dialogue Act	occurrences per piece placed
instruct	0.98
extend	2.39
repair	0.53
confirm	1.54

Table 2: *Dialogue act distributions in the Instruction Giver's speech*

gers action to start a new sub-goal (i.e. to place a new piece), showing that the IF has some initiative, using their knowledge of the task to continue un-instructed. On average there are over 2 additional extensions to the original utterance per sub-goal. Repairs occur on average in half of all sub-goals, though there is often more than one confirmation per piece placed, often because they refer to the action-level of taking the right piece as an intermediate confirmation rather than confirming the success of the entire sub-goal.

6. Conclusion

We have presented a proposal for an NLU method for simple pick-and-place robots which maps utterances to different levels in a task hierarchy. This attempts to ameliorate the surprising lack of overt imperative verb forms in natural task-oriented dialogue that we show in human-human data. This proposal shifts the task away from mapping utterances to either actions or goals exclusively, instead allowing flexible mapping to both levels during the interaction.

We also show how a continuous communicative grounding mechanism is vital for achieving fluid interaction and how confirmations and repairs can refer to both the goal and action levels. Reliance on overt conventionalized forms for achieving communicative grounding is limiting, and again NLU should estimate the mutual knowledge of the task structure when interpreting user speech.

We are currently developing a system incorporating this proposal, which will be fully evaluated with users.

7. Acknowledgments

We thank the anonymous reviewers for their helpful comments. This work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, funded by the German Research Foundation (DFG).

8. References

[1] H. H. Clark and S. E. Brennan, "Grounding in communication," *Perspectives on socially shared cognition*, vol. 13, no. 1991, 1991.

[2] D. L. Chen and R. J. Mooney, "Learning to interpret natural language navigation instructions from observations." in *AAAI*, vol. 2, 2011, pp. 1–2.

[3] C. Liu, S. Yang, S. Saba-Sadiya, N. Shukla, Y. He, S.-c. Zhu, and J. Chai, "Jointly learning grounded task structures from language instruction and visual demonstration," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, November 2016, pp. 1482–1492. [Online]. Available: https://aclweb.org/anthology/D16-1155

[4] J. Hough and D. Schlangen, "It's Not What You Do, It's How You Do It: Grounding Uncertainty for a Simple Robot," in *Proceedings of the 2017 Conference on Human-Robot Interaction (HRI2017)*, 2017.

[5] D. Schlangen and G. Skantze, "A General, Abstract Model of Incremental Dialogue Processing," *Dialoge & Discourse*, vol. 2, no. 1, 2011.

[6] S. Zarrieß, J. Hough, C. Kennington, R. Manuvinakurike, D. DeVault, R. Fernández, and D. Schlangen, "Pentoref: A corpus of spoken references in task-oriented dialogues," in *10th edition of the Language Resources and Evaluation Conference*, 2016.