Grounding Language by Continuous Observation of Instruction Following

Ting Han and **David Schlangen**
CITEC, Dialogue Systems Group, Bielefeld University
first.last@uni-bielefeld.de

Abstract

Grounded semantics is typically learnt from utterance-level meaning representations (e.g., successful database retrievals, denoted objects in images, moves in a game). We explore learning word and utterance meanings by continuous observation of the actions of an instruction follower (IF). While an instruction giver (IG) provided a verbal description of a configuration of objects, IF recreated it using a GUI. Aligning these GUI actions to sub-utterance chunks allows a simple maximum entropy model to associate them as chunk meaning better than just providing it with the utterance-final configuration. This shows that semantics useful for incremental (word-by-word) application, as required in natural dialogue, might also be better acquired from incremental settings.

1 Introduction

Situated instruction giving and following is a good setting for language learning, as it allows for the association of language with externalised meaning. For example, the reaction of drawing a circle on the top left of a canvas provides a visible signal of the comprehension of "*top left, a circle*". That such signals are also useful for machine learning of meaning has been shown by some recent work (*inter alia* (Chen and Mooney, 2011; Wang et al., 2016)). While in that work instructions were presented as text and the comprehension signals (goal configurations or successful navigations) were aligned with full instructions, we explore signals that are aligned more fine-grainedly, possibly to sub-utterance chunks of material. This, we claim, is a setting that is more representative of situated interaction, where typically no strict turn taking between instruction giving and execution is observed.

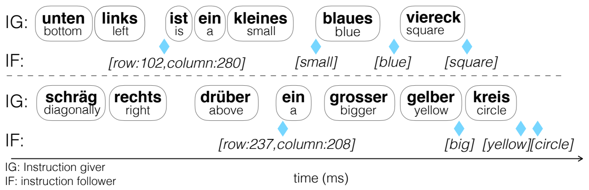


Figure 1: Example of collaborative scene drawing with IG words (rounded rectangles) and IF reactions (blue diamonds) on a time line.

Figure 1 shows two examples from our task. While the instruction giver (IG) is producing their utterance (in the actual experiment, this is coming from a recording), the instruction follower (IF) tries to execute it as soon as possible through actions in a GUI. The temporal placement of these actions relative to the words is indicated with blue diamonds in the figure. We use data of this type to learn alignments between actions and the words that trigger them, and show that the temporal alignment leads to a better model than just recording the utterance-final action sequence.

2 The learning task

We now describe the learning task formally. We aim to enable a computer to learn word and utterance meanings by observing human reactions in a scene drawing task. At the beginning, the computer knows nothing about the language. What it observes are an unfolding utterance from an IG and actions from an IF which are performed while the instruction is going on. Aligning each action a (or, more precisely, action *description*, as will become clear soon) to the nearest word w, we can represent an utterance / action sequence as follows:

$$w_{t_1}, w_{t_2}, a_{t_3}, \cdots, w_{t_i}, a_{t_{i+1}}, \cdots, w_{t_n} \qquad (1)$$

Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pages 491–496,
Valencia, Spain, April 3-7, 2017. ©2017 Association for Computational Linguistics

(Actions are aligned 'to the left', i.e. to the immediately preceding or overlapping word.)

As the IF concurrently follows the instruction and reacts, we make the simplifying assumption that each action a_{t_i} is a *re*action to the words which came before it and disregard the possibility that IF might act on a predictions of subsequent instructions. For instance, in (1), we assume that the action a_{t_3} is the interpretation of the words w_{t_1} and w_{t_2}. When no action follows a given word (e.g. w_{t_n} in (1)), we take this word as not contributing to the task.

We directly take these action symbols a as the representation of the utterance meaning so-far, or in other words, as its logical form; hence, the learning task is to predict an action symbol as soon as it is appropriate to do so. The input is presented as a chunk of the ongoing utterance containing at least the latest word. The utterance meaning U of a sequence $\{w_{t_1}, \ldots, w_{t_n}\}$ as a whole then is simply the concatenation of these actions:

$$U = \{a_{t_1}, ..., a_{t_i}\} \tag{2}$$

3 Modeling the learning task

3.1 Maximum entropy model

We trained a maximum entropy model to compute the probability distribution over actions from the action space $A = \{a^i : 1 \le i \le N\}$, given the current input chunk c:

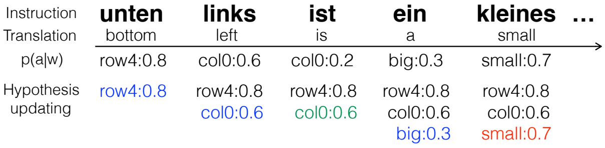
$$p(a^i|c) = \frac{1}{Z(c)} \exp \sum_j \lambda_j f_j(a^i, c) \tag{3}$$

λ_j is the parameter to be estimated. $f_j(a^i, c)$ is a simple feature function recording co-occurences of chunk and action:

$$f_j(a^i, c) = \begin{cases} 1 & \text{if } c = c_j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

In our experiments, we use a chunk size of 2, i.e., we use word bigrams. $Z(c)$ is the normalising constant. The logical form with the highest probability is taken to represent the meaning of the current chunk: $a^*(c) = \arg\max_i p(a^i|c)$

In the task that we test the approach on, the action space contains actions for locating an object in a scene; for sizing and colouring it; as well as for determining its shape. (See below.)



Instruction	unten	links	ist	ein	kleines ...
Translation	bottom	left	is	a	small
p(a|w)	row4:0.8	col0:0.6	col0:0.2	big:0.3	small:0.7
Hypothesis updating	row4:0.8	row4:0.8 col0:0.6	row4:0.8 col0:0.6	row4:0.8 col0:0.6 big:0.3	row4:0.8 col0:0.6 small:0.7

Figure 2: Example of hypothesis updating. New best hypotheses per type are shown in blue; retained hypotheses in green; revised hypotheses in red.

3.2 Composing utterance meaning

Since in our task each utterance places one object, we assume that each utterance hypothesis U contains a unique logical form for each of following concepts (referred as *type* of logical forms later): colour, shape, size, row and column position.

While the instruction unfolds, we update the utterance meaning hypotheses by adding new logical forms or updating the probabilities of current hypothesis. With each uttered word, we first check the type of the predicted logical form. If no logical form of the same type has already been hypothesised, we incorporate the new logical form to the current hypothesis. Otherwise, if the predicted logical form has a higher probability than the one with the same type in the current hypothesis, we update the hypothesis; if it has a lower probability, the hypothesis remains unchanged. Figure 2 shows an example of the hypothesis updating process.

4 Data collection

4.1 The experiment

While the general setup described above is one where IG gives an instruction, which a co-present IF follows concurrently, we separated these contributions for technical reasons: The instructions from IG were recorded in one session, and the actions from IF (in response to being played the recordings of IG) in another.

To elicit the instructions, we showed IGs a scene (as shown in the top of Figure 3) on a computer screen. They were instructed to describe the size, colour, shape and the spatial configuration of the objects. They were told that another person will listen to the descriptions and try to re-create the described scenes.

100 scenes were generated for the description task. Each scene includes 2 circles and a square. The position, size, colour and shape of each object were randomly selected when the scenes were

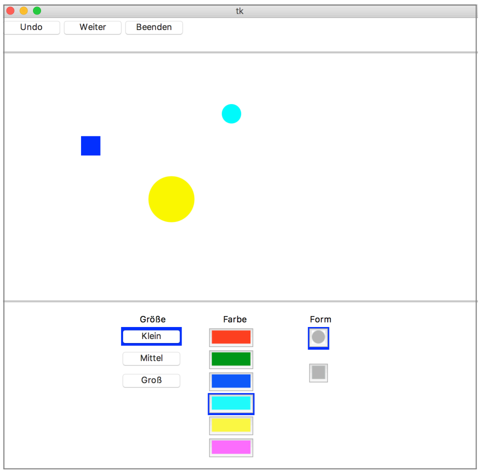


Figure 3: The GUI and a sample scene.

generated. The scenes were shown in the same order to all IGs. There was no time restriction for each description. Each IG was recorded for 20 minutes, yielding on average around 60 scene descriptions. Overall, 13 native German speakers participated in the experiment. Audio and video was recorded with a camera.

In the scene drawing task, we replayed the recordings to IFs who were not involved in the preceding experiment. To reduce the time pressure of concurrently following instructions and reacting with GUI operation, the recordings were cut into 3 separate object descriptions and replayed with a slower pace (at half the original speed). IFs decided when to begin the next object description, but were asked to act as fast as possible. This setup provides an approximation (albeit a crude one) to realistic interactive instruction giving, where feedback actions control the pace (Clark, 1996).

The drawing task was performed with a GUI (Figure 3) with separate interface elements corresponding to the aspects of the task (placing, sizing, colouring, determining shape). Before the experiment, IFs were instructed in using the GUI and tried several drawing tasks. After getting familiar with the GUI, the recordings were started. Overall, 3 native German speakers took part in this experiment. Each of them listened to the complete recordings of between 4 and 5 IGs, that is, to between 240 and 300 descriptions. The GUI actions were logged and timestamped.

4.2 Data preprocessing

Aligning words and actions First, the instruction recordings were manually transcribed. A forced-alignment approach was applied to temporally align the transcriptions with the recordings. Then, the IF actions were aligned with the recordings via logged timestamps.

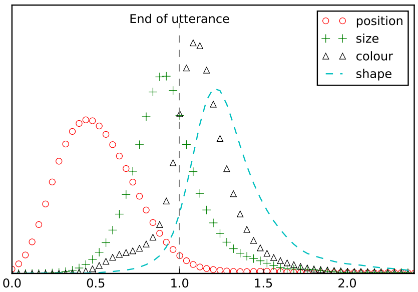


Figure 4: Action type distributions over utterance duration (1 = 100% of utterance played).

Figure 4 shows how action types distribute over utterances. As this shows, object positions tend to be decided on early during the utterance, with the other types clustering at the end or even after completion of the utterance.

Actions and Action Descriptions We defined a set of action symbols to serve as logical forms representing utterance chunk meanings. As described above, we categorised these action symbols into 5 types (shown in Table 1). The symbols were used for associating logical forms to words, while the type of actions was used for updating the hypothesis of utterance meaning (as explained in Section 3.2).

We make the distinction between actions and action symbol (or action description), because we make use of the fact that the same action may be described in different ways. E.g., a placing action can be described relative to the canvas as a whole (e.g. “bottom left”) or relative to other objects (e.g. “right of object 1”). We divided the canvas into a grid with 6 × 6 cells. We represent canvas positions with the grid coordinate. For example, `row1` indicates an object is in the first row of the canvas grid. We represent the relative positions with the subtraction of their indexes to corresponding referential objects. For example, `prev1_row1` indicates that the object is 1 row above the first described object. Describing the same action in these different ways gives us the required targets for associating with the different possible types of locating expressions.

Labelling words with logical forms With the assumption that each action is a *r*eaction to at most N words that came before it ($N = 3$ in our setup),

type	logical form
row	row1, row2 ... row6 prev1_row1, prev1_row2 ... prev2_row1, prev2_row2 ...
column	col1, col2 ... col6 prev1_col1, prev1_col2 ... prev2_col1, prev2_col2 ...
size	small, medium, big
colour	red, green, blue, magenta cyan, yellow
shape	circle, square

Table 1: Reaction types and logical forms.

we label these N previous words with the logical form of the action. E.g., for the first utterance from Figure 1 above:

	unten	links	ist	ein	kleines	blaues	Viereck
(1)	row4 col0	row4 col0	small	small blue	small blue square	blue square	square

Notice that a word might be aligned with more than one action, which means that the learning process has to deal with potentially noisy information. Alternatively, a word might not be aligned with any action.

5 Evaluation

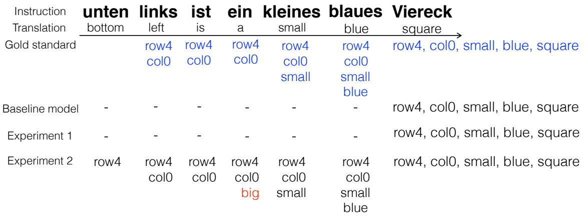
The data was randomly divided into train (80%) and test sets (20%). For our multi-class classification task, we calculated the F1-score and precision for each class and took the weighted sum as the final score.

Setup		F1-score	Precision	Recall
Proposed model	Exp1	**0.75**	**0.65**	**0.89**
	Exp2	0.66	0.55	0.83
Baseline model		0.60	0.52	0.71

Table 2: Evaluation results.

Figure 5 illustrates the evaluation process of each setup.

Proposed model The proposed model was evaluated on the utterance and the incremental level. In **Experiment 1**, the meaning representation is *assembled* incrementally as described above, but evaluated utterance-final. In **Experiment 2**, the model is evaluated incrementally, after each word of the utterance. Hence, late predictions (where a part of the utterance meaning is predicted later than would have been possible) are penalised in Experiment 2, but not Experiment 1. The model performs better on the utterance level, which suggests that the hypothesis updating process can successfully revise false interpretations while the descriptions unfold.



Instruction	unten	links	ist	ein	kleines	blaues	Viereck
Translation	bottom	left	is	a	small	blue	square
Gold standard		row4 col0	row4 col0	row4 col0	row4 col0 small	row4 col0 small blue	row4, col0, small, blue, square
Baseline model	-	-	-	-	-	-	row4, col0, small, blue, square
Experiment 1	-	-	-	-	-	-	row4, col0, small, blue, square
Experiment 2	row4	row4 col0	row4 col0	row4 col0 big	row4 col0 small	row4 col0 small blue	row4, col0, small, blue, square

Figure 5: Evaluation Setups. Exp. 1 only evaluates the utterance-final representation, Exp. 2 evaluates incrementally. False interpretations are shown in red.

Baseline model For comparison, we also trained a baseline model with temporally unaligned data (comparable to a situation where only at the end of an utterance a gold annotation is available). For (1), this would result in all words getting assigned the labels row4,col0,small,blue,square. As Table 2 shows, this model achieves lower results. This indicates that temporal alignment in the training data does indeed provide better information for learning.

Error analysis While the model achieves good performance in general, it performs less well on position words. For example, given the chunk "schräg rechts" (*diagonally to the right*) which describes a landmark-relative position, our model learned as best interpretation a canvas-relative position. The hope was that offering the model the two different action description types (canvas-relative and object-relative) would allow it to make this distinction, but it seems that here at least the more frequent use of "rechts" suppresses that meaning.

6 Related work

There has been some recent work on grounded semantics with ambiguous supervision. For example, Kate and Mooney (2007) and Kim and Mooney (2010) paired sentences with multiple representations, among which only one is correct. Börschinger et al. (2011) introduced an approach to ground language learning based on unsupervised PCFG induction. Kim and Mooney (2012) presents an enhancement of the PCFG approach that scales to such problems with highly-ambiguous supervision. Berant et al. (2013) and Dong and Lapata (2016) map natural language to machine interpretable logical forms with question-answer pairs. Tellex et al. (2012), Salvi et al. (2012), Matuszek et al. (2013), and Andreas and Klein (2015) proposed approaches to learn grounded semantics from natural language and action associations. These approaches paired ambiguous robot actions with natural language descriptions from humans. While these approaches achieve good learning performance, the ambiguous logical forms paired with the sentences were manually annotated. We attempted to align utterances and potential logical forms by continuously observing the instruction following actions. Our approach not only needs no human annotation or prior pairing of natural language and logical forms for the learning task, but also acquires less ambiguous language and action pairs. The results show that the temporal information helps to achieve competitive learning performance with a simple maximum entropy model.

Learning from observing successful interpretation has been studied in much recent work. Besides the work discussed above, Frank and Goodman (2012), Golland et al. (2010), and Reckman et al. (2010) focus on inferring word meanings through game playing. Branavan et al. (2009), Artzi and Zettlemoyer (2013), Kollar et al. (2014) and Monroe and Potts (2015) infer natural language meanings from successful instruction execution of humans/agents. While interpretations were provided on utterance level in above works, we attempt to learn word and utterance meanings by continuously observing interpretations of natural language in a situated setup which enables exploitation of temporally-aligned instruction giving and following.

7 Conclusions

Where most related work starts from utterance-final representations, we investigated the use of more temporally-aligned understanding data. We found that in our setting and for our simple learning methods, this indeed provides a better signal. It remains for future work to more clearly delineate the types of settings where such close alignment on the sub-utterance level might be observed.

Acknowledgments

This work was supported by the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277) at Bielefeld University, which is funded by the German Research Foundation (DFG). The first author would like to acknowledge the support from the China Scholarship Council.

References

Jacob Andreas and Dan Klein. 2015. Alignment-based compositional semantics for instruction following. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1165–1174*, pages 1165–1174, Lisbon, Portugal. Association for Computational Linguistic.

Yoav Artzi and Luke Zettlemoyer. 2013. Weakly supervised learning of semantic parsers for mapping instructions to actions. *Transactions of the Association for Computational Linguistics*, 1:49–62.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, volume 2, page 6.

Benjamin Börschinger, Bevan K. Jones, and Mark Johnson. 2011. Reducing grounded learning tasks to grammatical inference. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1416–1425. Association for Computational Linguistics.

Satchuthananthavale R.K. Branavan, Harr Chen, Luke S. Zettlemoyer, and Regina Barzilay. 2009. Reinforcement learning for mapping instructions to actions. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, volume 1, pages 82–90. Association for Computational Linguistics.

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, San Francisco, California. AAAI Press.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge.

Li Dong and Mirella Lapata. 2016. Language to logical form with neural attention. In *The 54th Annual Meeting of the Association for Computational Linguistics*, pages 2368–2378, Berlin, Germany. Association for Computational Linguistics.

Michael C. Frank and Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science*, 336(6084):998–998.

Dave Golland, Percy Liang, and Dan Klein. 2010. A game-theoretic approach to generating spatial descriptions. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 410–419. Association for Computational Linguistics.

Rohit J. Kate and Raymond J. Mooney. 2007. Learning language semantics from ambiguous supervision. In *AAAI*, volume 7, pages 895–900.

Joohyun Kim and Raymond J. Mooney. 2010. Generative alignment and semantic parsing for learning from ambiguous supervision. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 543–551. Association for Computational Linguistics.

Joohyun Kim and Raymond J. Mooney. 2012. Unsupervised pcfg induction for grounded language learning with highly ambiguous supervision. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 433–444. Association for Computational Linguistics.

Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. 2014. Grounding verbs of motion in natural language commands to robots. In *Experimental robotics*, pages 31–47. Springer.

Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. 2013. Learning to parse natural language commands to a robot control system. In *Experimental Robotics*, pages 403–415. Springer.

Will Monroe and Christopher Potts. 2015. Learning in the rational speech acts model. In *In Proceedings of 20th Amsterdam Colloquium, Amsterdam, December. ILLC*.

Hilke Reckman, Jeff Orkin, and Deb K. Roy. 2010. Learning meanings of words and constructions, grounded in a virtual game. In *Proceedings of the Conference on Natural Language Processing 2010*, pages 67–75, Saarbrücken, Germany. Saarland University Press.

Giampiero Salvi, Luis Montesano, Alexandre Bernardino, and Jose Santos-Victor. 2012. Language bootstrapping: Learning word meanings from perception–action association. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 42(3):660–671.

Stefanie Tellex, Pratiksha Thaker, Josh Joseph, Matthew R. Walter, and Nicholas Roy. 2012. Toward learning perceptually grounded word meanings from unaligned parallel data. In *Proceedings of the Second Workshop on Semantic Interpretation in an Actionable Context*, pages 7–14. Association for Computational Linguistics.

Sida I. Wang, Percy Liang, and Christopher D. Manning. 2016. Learning language games through interaction. In *The 54th Annual Meeting of the Association for Computational Linguistics*, pages 2368–2378, Berlin, Germany. Association for Computational Linguistics.