KILLE: learning grounded language through interaction

Simon Dobnik and Erik Wouter de Graaf
Dept. of Philosophy, Linguistics & Theory of Science
University of Gothenburg, Sweden
simon.dobnik@gu.se and kille@masx.nl

Abstract

Testing and computational implementation of formal models of situated linguistic interaction imposes demands on computational infrastructure. We present our system called KILLE and provide a proof-of-concept evaluation of interactive situated learning of object categories and spatial relations.

1 Grounded meaning in interaction

Contemporary approaches to semantics of natural language (Cooper, 2016; Fernández et al., 2011) are based on two important premises: (i) meanings are not universal and static but are agent-relative and are continuously adapted in interaction with other agents and environment (Clark, 1996; Pickering and Garrod, 2004); and (ii) meanings (sense and reference) are multi-modal where different lexical items are sensitive to different modalities in different contexts to different degrees (Coventry and Garrod, 2005).

Both aspects have changed the focus in computational semantics from engineering formal rules that cover a domain or a fragment of linguistic data off-line to approaches that are data driven and involve continuous online fine-tuning of the model's parameters (Skočaj et al., 2011; Matuszek et al., 2012). In robotics a shift in the approach has happened much earlier as it quickly became apparent that robots with static models cannot deal with any changes in the environment or with the environment's uncertainty. Instead, modern robotics uses models which are learned from data and refined continuously as the robot's interaction with the environment develops (for example (Dissanayake et al., 2001) for map building). We argue that the same paradigm should also be adopted when dealing with computational models of language. In this view the focus of building a computational system is not on designing representations but investigating and modelling interactive strategies or dialogue games (Kowtko et al., 1992) that will allow construction of such representations or fine-tuning of their features, depending on how much of representations are pre-available to such a system.[1]

The interactive semantics of a computational system have also implications on the models of meaning used. The predominant semantic representations used in computational semantics today are vector-space representations that define meaning as semantic similarity between lexical items on the basis of their co-occurrence in contexts (Turney et al., 2010; Clark, 2015). Such models can be successfully extracted from large corpora of text and are very successful in representing meaning. However, they nonetheless represent meaning in an indirect way as they never consider a relation between an expression and situations in which that expression applies to or is true for. The reason why words in particular linguistic contexts are lexically similar is because words in linguistic strings as a whole refer to (more or less) the same situations which we do not have access to or ignore when we built vector space models. However, in an interactive scenario described above we can explore linking linguistic expressions and perceptual features directly, a process which is commonly known as grounding (Harnad, 1990; Roy, 2002). Such models are required for situated dialogue agents or conversational robots which have to link language and situations that they jointly attend to with human conversational partners.[2]

[1] This sounds similar to the Chomsky's innateness claim but here we are thinking of purely engineering a system and make no claims about human cognition.

[2] It is important to emphasise nonetheless that vector space models may provide an important source of back-

Grounded meanings of linguistic descriptions such as “close to the table” and “red” correspond to some function from physical or colour space to a degree of acceptability of that description (Logan and Sadler, 1996; Roy, 2002; Skočaj et al., 2011; Matuszek et al., 2012; Kennington and Schlangen, 2015; McMahan and Stone, 2015). Cognitive structures are hierarchically organised at several representation layers focusing on and combining different modalities (Kruijff et al., 2007). Since the functions predict distributions of degree of applicability several descriptions may be equally applicable for the same perceptual situation: the chair can be “close to the table” or “to the left of the table” which means *vagueness* is prevalent in grounding. This however, can be resolved through interaction by adopting appropriate interaction strategies (Kelleher et al., 2005; Skantze et al., 2014; Dobnik et al., 2015).

A formal model of perceptual semantics in interaction has been the focus of Type Theory with Records (TTR) (Cooper, 2016; Larsson, 2013; Dobnik et al., 2013). Implementing, validating and testing such models imposes complex demands on computational infrastructure in the sense that this involves connecting perceptual sensors with dialogue systems and machine learning algorithms. Processing language in interaction also presents challenges from the computational perspective as it is often not trivial to employ existing language technology tools and (machine learning) algorithms, which were developed for processing data offline, in an interactive tutoring scenario. To address both issues we have developed a framework for situated agents that learn grounded language incrementally and online with a help of human tutor called KILLE[3] (Kinect Is Learning LanguagE). This paper focuses on the construction of the Kille framework and its properties while it also provides a proof-of-concept evaluation of such learning of simple object and spatial relations representations. We hope that this framework will be a useful tool for future studying and computational modelling language in interaction.

2 The KILLE system

KILLE is a non-mobile table-top robot connecting Kinect sensors with image processing (*libfreenect*), classification (clustering of visual features and location classification) and a spoken dialogue system OpenDial[4] (Lison, 2013) connected through Robot Operating System (ROS) (Quigley et al., 2009). The latter is a popular robotic middle-ware which ensures communication between them. It runs on a variety of popular robotic hardware implementations which means that our system could be ported to them without too much modification (Figure 1). We prefer a robotic middle-ware rather systems centred around dialogue systems because it allows us to represent and exchange perceptual and linguistic information together and in the same way: there is one information state for both. In addition to the integration of these modules, our main contribution is implementation of ROSDial which provides and interface between OpenDial and ROS, implementation of Kille Core which provides perceptual and spatial classification, and implementation of dialogue games that interface between dialogue and perceptual classification and therefore enable incremental perceptual learning.

The system learns to recognise objects presented to it by a human tutor from scratch. It can direct learning by asking for more objects of a particular category if it is not able to classify them with sufficient reliability, thus filling in the missing knowledge. If more objects of a particular category are available in the scene and the system is able to recognise them, the system queries the user to describe spatial relations between them. Each of these kinds of descriptions focus on a different perceptual features (visual features vs location) and represent two fundamental linguistic semantic categories: entities and entity relations. Overall, KILLE combines both passive (being instructed) and active (requesting examples to learn from) learning which is incremental in terms of learning to recognise situations denoted by the lexical categories. A demo of the system can be found at `http://www.youtube.com/watch?v=AsEgaka6tH0`.

Below there are some examples of dialogue games that we implemented as dialogue rules in the dialogue manager which act upon both linguistic and perceptual knowledge.

ground knowledge in such scenario and hence a dialogue agent does not have to learn every meaning representation through grounding. The challenges of integration of both meaning representations are a focus of ongoing research.

[3] Swedish for “fellow”, “chap” or “bloke”.

[4] https://github.com/plison/opendial

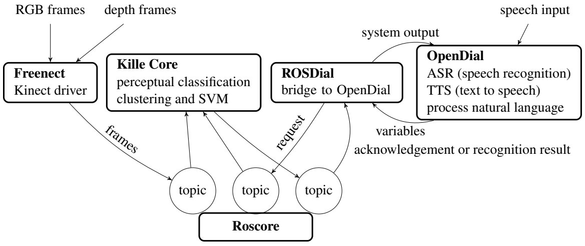

Figure 1: Kille modules

1. **Direct instruction: objects**

   User: This is a cup.
   System: [Object is focused on and observed.]
   S: OK, I learned a cup.

2. **Describing objects and relations with (un)certainty**

   U: What is this?
   S: [Object is focused on and observed.]
   U: The object is thought to be a book, but it might also be a mug.

3. **Reinforcement from feedback**

   U: What is this?
   S: [Object is focused on and observed.]
   S: This is a cup.
   U: That is correct.
   S: [Observation is added to the models.]

4. **Self initiative: spatial relations**

   S: [Objects are focused on and observed.]
   S: One of the objects is book, the other one is box. Please tell me where things are.
   U: The book is to the left of the box.
   S: OK, this relation is called “to the left of”.

For visual representations we use Open Source Computer Vision (OpenCV)[5] (Bradski and Kaehler, 2008) which is a popular library for computer vision including real-time machine learning applications. Through ROS we receive real-time frames from Kinect which includes both data from the depth sensor and the visual RGB sensor. We use the depth information (which gives us a precise 3d location of points making up an object) to detect the object in focus and later take the pixels representing these points in focus to detect SIFT features (Scale-Invariant Feature Transform) (Lowe, 1999) over them which are used to represent objects in our model as shown in Figure 2.

Objects, including those that are very similar and belong to the same category, have different number of SIFT descriptors detected depending on their visual properties: some objects have more visual details than others. There is a bias that object with less features match objects with more (and similar looking) features. In our interactive scenario there is also no guarantee that the same features will be detected after the object is reintroduced (or even between two successive scans) as the captured frame will be slightly different from the previously captured one because of slight changes in location, lighting and camera noise.

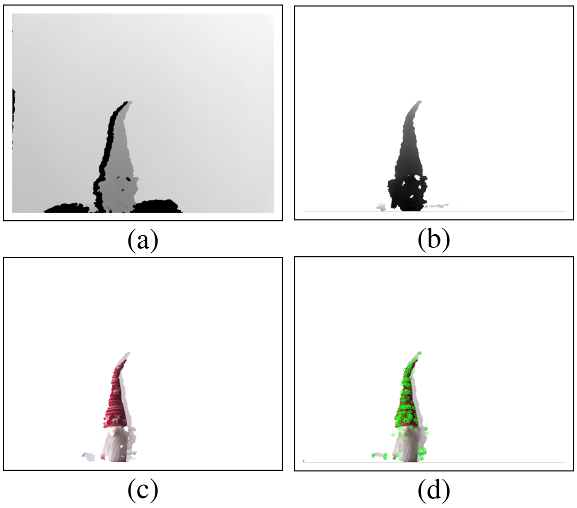

Figure 2: A perception of a plush gnome from the depth sensor (a) including the background, (b) with the background removed, (c) with the RGB image superimposed, and (d) with SIFT features detected in the image. The black border in (a) is a perceptual artefact arising from the interference of sensors.

3 Interactive perceptional learning

In the following subsections we present a proof-of concept implementation and evaluation of perceptual learning through interaction which demonstrates the usability of the Kille framework.

Learning to recognise objects To recognise objects we developed a nearest neighbour classification method based on the the FLANN library (Muja and Lowe, 2009) which works by comparing the SIFT descriptors of object to classify with the objects in the database and then returns the class of the closest matching object. In the evaluation, 10 consecutive scans are taken and their recognition scores are averaged to a single score. This improves the accuracy but increases the classification time (which is nonetheless still reasonable for the small domain of objects we are con-

[5] http://opencv.org

sidering). The location of the recognised object is estimated by taking the locations of the twenty matched descriptors with the shortest distance.

To evaluate the system's performance in an interactive tutoring scenario we chose the following 10 objects: apple, banana, teddy bear, book, cap, car, cup, can of paint, shoe and shoe-box. A human tutor successively re-introduces the same 10 objects to the system in a pre-defined order over four rounds trying to keep the presentation identical as much as possible. In each round all objects are first learned and then queried. To avoid ASR errors both in learning and generation text input is used.

Taking the average SIFT feature matching scores over 4 rounds for each object and taking the class of the object with highest mean score, on average all but one object were recognised correctly. However, the cap was consistently confused with the banana. There were a couple of individual confusions that have been levelled out in the calculation of the average score. To test how distinct objects are from one another we calculated a difference of the matching scores of the highest-ranking object of the correct category and the other highest ranking candidate. If we arrange objects by this score, we get the following ranking (from more distinct to least distinct): book > car > shoe > cup > banana > bear > apple > paint > shoe-box > cap. We also tested recognition of the same objects when rotated and recognition of new objects of the same category.

Learning to recognise spatial relations Before spatial relations can be learned the system must recognise the target and the landmark objects ("the gnome/TARGET is to the left of the book/LANDMARK") both in a linguistic string and in a perceptual scene. Twenty highest ranking SIFT features are taken for each object and their x (width), y (height) and z (depth) coordinates are averaged, thus giving us the centroid of the 20 most salient features of an object. The coordinate frame of the coordinates is transposed to the centre of the landmark object. The relativised location of the target to the landmark are fed to a Linear Support Vector Classifier (SVC) with descriptions as target classes.

A human tutor taught the system by presenting it the target object (a book) randomly 3 times at 16 different locations (2 distances/circles containing 8 points separated at 45°) in relation to the landmark (the car). The spatial descriptions that the human instructor used were *to the left of*, *to the right of*, *in front of*, *behind of*, *near* and *close to* (6). The performance of the system was evaluated by two human conversational partners, one of whom was also the tutor from the learning stage. The target object was randomly placed in one of the 16 locations and each location was used twice which gave us 32 generations. A particular location may be described with several spatial descriptions (but not all combinations of descriptions are possible) but some may be more appropriate than others. The evaluators first wrote down a description they would use to describe the scene and then the system would be queried about the location of the target to which it provided a response. The evaluators would then also record whether they agree with the generation. The observed blind agreement between the evaluators is 0.5313 with $\kappa = 0.4313$ which means that choosing a spatial description is quite a subjective task. The blind agreement between the evaluators and the system is 0.2344 with $\kappa = 0.0537$. The evaluators were happy with the system's generation in additional 37.5% of cases, which means that the system generated an appropriate description in 60.94% of cases which is encouraging and comparable to the similar task in the literature. Note also that the system tried to learn continuous functions from a very small number of examples, on an average only 46/6=8 instances.

4 Conclusion and future work

In this paper we argue that there is a need for a computational infrastructure that will allow us modelling dynamic grounded semantics in interaction for two reasons: (i) to verify semantic theories and (ii) to provide a platform for their computational implementations. We developed and framework called KILLE a simple interactive "robot" which we argue provides a good solution for modelling these aspects and at the same time can be ported to more sophisticated robotic hardware platforms. We demonstrated a proof-of-concept of learning object categories and spatial relations following the theoretical proposals in the literature. We hope that the platform will provide useful for testing further models of linguistic and perceptual interactions.

References

Gary Bradski and Adrian Kaehler. 2008. *Learning OpenCV: Computer vision with the OpenCV library*. O'Reilly Media, Inc.

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

Stephen Clark. 2015. Vector space models of lexical meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantics — second edition*, chapter 16, pages 493–522. Wiley – Blackwell.

Robin Cooper. 2016. Type theory and language: From perception to linguistic communication. Draft of chapters 1-6, 30th November.

Kenny Coventry and Simon Garrod. 2005. Spatial prepositions and the functional geometric framework. towards a classification of extra-geometric influences. In Laura Anne Carlson and Emile van der Zee, editors, *Functional features in language and space: insights from perception, categorization, and development*, volume 2, pages 149–162. OUP.

M. W. M. G Dissanayake, P. M. Newman, H. F. Durrant-Whyte, S. Clark, and M. Csorba. 2001. A solution to the simultaneous localization and map building (SLAM) problem. *IEEE Transactions on Robotic and Automation*, 17(3):229–241.

Simon Dobnik, Robin Cooper, and Staffan Larsson. 2013. Modelling language, action, and perception in Type Theory with Records. In Denys Duchier and Yannick Parmentier, editors, *Constraint Solving and Language Processing (CSLP 2012), Revised Selected Papers*, v8114 of *LNCS*, pages 70–91. Springer Berlin Heidelberg.

Simon Dobnik, Christine Howes, and John D. Kelleher. 2015. Changing perspective: Local alignment of reference frames in dialogue. In *Proceedings of goDIAL - Semdial 2015*, pages 24–32, Gothenburg, Sweden, 24–26th August.

Raquel Fernández, Staffan Larsson, Robin Cooper, Jonathan Ginzburg, and David Schlangen. 2011. Reciprocal learning via dialogue interaction: Challenges and prospects. In *Proceedings of the IJCAI 2011 ALIHT Workshop*, Barcelona, Catalonia, Spain.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D*, 42(1–3):335–346, June.

J.D. Kelleher, F. Costello, and J. van Genabith. 2005. Dynamically structuring updating and interrelating representations of visual and linguistic discourse. *Artificial Intelligence*, 167:62–102.

Casey Kennington and David Schlangen. 2015. Simple learning and compositional application of perceptually grounded word meanings for incremental reference resolution. In *ACL-IJCNLP 2015*, pages 292–301, Beijing, China, July. ACL.

Jacqueline C Kowtko, Stephen D Isard, and Gwyneth M Doherty. 1992. Conversational games within dialogue. HCRC research paper RP-31, University of Edinburgh.

Geert-Jan M. Kruijff, Hendrik Zender, Patric Jensfelt, and Henrik I. Christensen. 2007. Situated dialogue and spatial organization: what, where... and why? *International Journal of Advanced Robotic Systems*, 4(1):125–138. Special issue on human and robot interactive communication.

Staffan Larsson. 2013. Formal semantics for perceptual classification. *Journal of Logic and Computation*, online:1–35, December 18.

Pierre Lison. 2013. *Structured Probabilistic Modelling for Dialogue Management*. Ph.D. thesis, Department of Informatics, Faculty of Mathematics and Natural Sciences, University of Oslo, 30th October.

Gordon D. Logan and Daniel D. Sadler. 1996. A computational analysis of the apprehension of spatial relations. In Paul Bloom, Mary A. Peterson, Lynn Nadel, and Merrill F. Garrett, editors, *Language and Space*, pages 493–530. MIT Press, Cambridge, MA.

David G Lowe. 1999. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. IEEE.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of ICML 2012*, Edinburgh, Scotland, June 27th - July 3rd.

Brian McMahan and Matthew Stone. 2015. A Bayesian model of grounded color semantics. *Transactions of the ACL*, 3:103–115.

Marius Muja and David G Lowe. 2009. Fast approximate nearest neighbors with automatic algorithm configuration. *VISAPP (1)*, 2(331–340):2.

Martin J. Pickering and Simon Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Y Ng. 2009. ROS: an open-source robot operating system. In *ICRA workshop on open source software*, volume 3, page 5.

Deb K. Roy. 2002. Learning visually-grounded words and syntax for a scene description task. *Computer speech and language*, 16(3):353–385.

Gabriel Skantze, Anna Hjalmarsson, and Catharine Oertel. 2014. Turn-taking, feedback and joint attention in situated human-robot interaction. *Speech Communication*, 65:50–66.

Danijel Skočaj, Matej Kristan, Alen Vrečko, Marko Mahnič, Miroslav Janíček, Geert-Jan M. Kruijff, Marc Hanheide, Nick Hawes, Thomas Keller, Michael Zillich, and Kai Zhou. 2011. A system for interactive learning in dialogue with a tutor. In *IROS 2011*, San Francisco, CA, USA, 25-30 September.

Peter D Turney, Patrick Pantel, et al. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.