Structured Knowledge and Kernel-based Learning: the case of Grounded Spoken Language Learning in Interactive Robotics

Roberto Basili and Danilo Croce

Department of Enterprise Engineering
University of Roma, Tor Vergata
{basili,croce}@info.uniroma2.it

Abstract. Recent results achieved by statistical approaches involving Deep Neural Learning architectures suggest that semantic inference tasks can be solved by adopting complex neural architectures and advanced optimization techniques. This is achieved even by simplifying the representation of the targeted phenomena. The idea that representation of structured knowledge is essential to reliable and accurate semantic inferences seems to be implicitly denied. However, Neural Networks (NNs) underlying such methods rely on complex and beneficial representational choices for the input to the network (e.g., in the so-called pre-Training stages) and sophisticated design choices regarding the NNS inner structure are still required.

While optimization carries strong mathematical tools that are crucially useful, in this work, we wonder about the role of representation of information and knowledge. In particular, we claim that representation is still a major issue, and discuss it in the light of Spoken Language capabilities required by a robotic system in the domain of service robotics. The result is that adequate knowledge representation is quite central for learning machines in real applications. Moreover, learning mechanisms able to properly characterize it, through expressive mathematical abstractions (i.e. trees, graphs or sets), constitute a core research direction towards robust, adaptive and increasingly autonomous AI systems.

Recent results achieved by statistical approaches involving Deep Neural Learning architectures (as in [1]) suggest that semantic inference tasks can be solved by adopting complex neural architectures and advanced mathematical optimization techniques, but simplifying the representation of the targeted phenomena. The idea that representation of structured knowledge is essential to reliable and accurate semantic inferences seems to be implicitly denied. As an example, the application of Deep Neural Networks architectures in the context of Natural Language Processing or Machine Translation is quite radical in this respect, since the work presented in [2].

However, Neural Networks (NNs) underlying such methods rely on beneficial representational choices for the input to the network (e.g., in the so-called pre-training stages) and complex design choices regarding the NNs inner structure are still required ([3]). Moreover, some recent works suggest that optimal hyper-parameterization of huge networks is possible, thus making the differences between different architectures even less relevant (as discussed for example in [4]). While optimization carries strong mathematical tools that are crucially useful, in this work, we wonder here about the role of representation of information and knowledge.

A large body of research on the integration of background knowledge with the learning algorithms has been early carried out within the framework of Inductive Logic Programming (ILP), presented in [5]. ILP is useful for logically encoding background knowledge and extensions to standard ILP algorithms have been proposed for encoding syntactic and semantic relational information of a knowledge base in the kernel function, thus providing a unified, flexible treatment of structured and non-structured data. More recently, in [6], the induction of set of clauses in a First Order Inductive Learner has been integrated and used as features in standard kernel methods. In this way, principled, theory-driven data representations result in kernels that allow consistent inferences in SVM-based classification and regression tasks.

In this work, we claim that representation is still a major issue, and discuss it in the light of Spoken Language capabilities required by a robotic system in the domain of service robotics. End-to-end communication processes in natural language are challenging for robots for the deep interaction of different cognitive abilities. For a robot to react to a user command like "*Take the book on the table*" a number of implicit assumptions should be met to understand its possibly ambiguous content. First, at least two entities, a `book` and a `table`, must exist in the environment and the speaker expects the robot to be aware of such entities. Accordingly, the robot must have access to an inner representation of the objects, e.g. an explicit map of the environment. Second, mappings from words, i.e. lexical references, to real world entities must be available. *Grounding* here [7] links symbols (e.g. words) to the corresponding perceptual information.

Spoken Language Understanding (SLU) in interactive dialogue systems acquires a specific nature, when applied in Interactive Robotics. Linguistic interactions are context aware in the sense that both the user and the robot access and make reference to the environment (i.e. entities of the real world). In the above example, "*taking*" is the intended action whenever a book is actually on the table, so that *the book on the table* refers to a unique semantic role, i.e. to one entity playing an explicit role in the command (that is "*the book to be taken actually located on a table*"). On the contrary, the command may refer to a "*bringing*" action, when no book is on the table and *the book* and *on the table* correspond to different roles. Robot interactions need thus to be *grounded*, as meaning must correspond to the physical world and interpretation is strongly interlaced with what is perceived, as pointed out by psycho-linguistic theories [8]. As a consequence, a correct interpretation is more than a linguistically motivated mapping from an audio signal (e.g. the spoken command) to a meaning representation formalism compatible with a linguistic theory (e.g., semantic frames as discussed in [9]). Correctness implies also physical coherence, as entities in the environment must be known and the intended predicates must correspond to (possibly known) actions coherent with the environment, too.

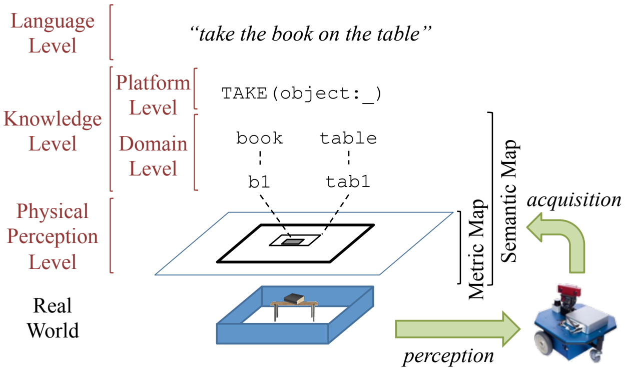

Fig. 1. Levels of representation in interactive robotics

While traditional SLU mostly relies on linguistic information contained in texts (i.e., derived only from transcribed words), its application in Interactive Robotics depends on a variety of other factors, including the perception of the environment.

We can organize these factors into a layered representation as shown in Figure 1. First, we rely on the *language level* that governs linguistic inferences: it includes observations (e.g. sequences of transcribed words) as well as the linguistic assumptions of the speaker, here modeled through frame-like predicates by which the inner lexicon can be organized. Similarly, evidences involved by the robot's perception of the world must be taken into account. The physical level, i.e. the real world, is embodied in the *physical perception level*: we assume that the robot has an image of this world where the existence and the spatial properties of entities are represented. Such representation is built by mapping the direct input of robot sensors into geometrical representations, e.g. *metric maps*. These provide a structure suitable for anchoring the *knowledge level*. Here *symbols* (i.e., knowledge primitives) are used to refer to real world entities and their properties inside the *domain level*. This comprises all active concepts the robot is aware of, as they are realized in a specific environment, that refer to general knowledge (e.g. conceptual categories) it has about the domain. All this information plays a crucial role during linguistic interactions. The integration of topological, i.e. metric, information with notions related to the knowledge level provides an augmented representation of the environment, called *semantic map* [10]. In this map, the existence of real world objects can be associated to *lexical* information, in the form of entity names given by a knowledge engineer or spoken by a user while pointing to an object, as in Human-Augmented Mapping [11,12]. It is worth noticing that the robot itself is a special entity described at this knowledge level: it does know its constituent parts as well as its capabilities, that are the actions it is able to perform. In our case, we introduce an additional level (namely *platform level*), whose information is instantiated in a knowledge base called *Platform Model*. In this way, a comprehensive perceptual knowledge level is made available, including both a model of the world and a model of the robot itself.

While SLU for Interactive Robotics have been mostly carried out over the evidences specific to the linguistic level, e.g., in [13,14,15], we argue that such process should deal with all the aforementioned layers in an harmonized and coherent manner. All linguistic primitives, including predicates and semantic arguments, correspond to perceptual counterparts, such as plans, robot's actions or entities involved in the underlying events.

The positive impact of such layers of knowledge in the automatic interpretation of robotic commands expressed in natural language has been presented in [16], where the final interpretation process depends not only on the linguistic information, but also on the perceptual knowledge level. This process is expected to produce interpretations that coherently mediate among the world (with all the entities composing it), the robotic platform (with all its inner representations and its capabilities) and the pure linguistic level triggered by a sentence. To this end, a discriminative approach[1] to SLU has been adopted. Grounded information is here directly injected within a structured learning algorithm, that is SVM-HMM [17]. Such integration of linguistic and perceptual knowledge significantly improves the quality and robustness of the overall interpretation process, as up to a 38% of reduction in the relative error is observed ([16]). Integration is achieved by feeding the learning algorithms with a representation where perceptual knowledge extracted from a semantic map is made available through explicit features: it derives from a grounding mechanism based on the evidences triggered by linguistic references and distributional semantic similarity. Moreover, SVM classification based on multiple kernels is adopted to integrate the different features.

Kernels introduce a second crucial issue in the role of representations in machine learning method, in particular those applied to complex decision functions: the readability of the resulting models. Understanding *why* a data-driven method provided a specific answer will be crucial as this model will be integrated in everyday life. This issue has been for example faced in the task of Automatic Generation of Image caption [3]: an Attention-based model has been there used to extend a Deep Learning architecture and focus on the image portion that stimulated the generation of a particular caption. In semantic inference tasks involving natural language, it will be crucial to understand the reason a text triggered a particular output of a data-driven method: for example, in sentiment analysis over Twitter, we should know which words in an input tweet are responsible to evoke the output sentiment class. We foster here the importance of kernel methods [18]. They allow the application of learning algorithms over discrete structures that directly express the linguistic information underlying

[1] This method is implemented in the adaptive spoken Language Understanding For Robots (LU4R) processing chain: `http://sag.art.uniroma2.it/lu4r.html`

input texts. As an example, the adoption of Tree Kernels [19] methods allows to directly apply machine learning methods, such as Support Vector Machines, over tree structures that are directly produced by a Syntactic Parser. The cognitive role of trees in most syntactic theories implies that tree-kernel based feature engineering is most closely related to human-like language learning and suggest more natural generalization processes.

Moreover, the model underlying the decision function for this class of methods, for example a classifier recognizing the target of a question in natural language or the semantic role to be assigned to the argument of a linguistic predicate ([20,21]), depends only an a subset of training examples, the core that is *crucial* for the final decision. The learning algorithm (e.g. a batch SVM) just assigns non-zero weights to only those training examples at the frontier (i.e. the so-called support vectors): these are the only ones that contribute to the final decision. Notice that for this class of leaning algorithms, examples are directly selected by the learning algorithm. They can be expected to reflect the implicit linguistic knowledge used by the speaker to decide. The linguistic structures corresponding to such selected core example set, i.e. the trees or subtrees corresponding to the support vectors, provide important information to increase the readability of the system behavior. Kernels corresponds thus to a straightforward learning method where a good trade-off between readability and accuracy is quite naturally achieved.

In synthesis, adequate representations for many different aspects of human knowledge appears still quite central for learning machines in real applications. Although these give rise to very powerful and accurate inferences regarding uncertain decisions, they are usually the side effects of complex design choices regarding the task and the input representation: these are all but ontological assumptions about the inference process and the background world model. However, complex learning mechanisms able to properly characterize the different representational properties through expressive mathematical models (i.e. trees, graphs or sets as in convolution kernels) exist and have already successfully applied. They constitute a core research direction towards robust, adaptive and increasingly autonomous AI systems whose models are readable and increasingly expressive.

References

1. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553) (05 2015) 436–444
2. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. J. Mach. Learn. Res. **12** (November 2011) 2493–2537
3. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A.C., Salakhutdinov, R., Zemel, R.S., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. CoRR **abs/1502.03044** (2015)
4. Levy, O., Goldberg, Y., Dagan, I.: Improving distributional similarity with lessons learned from word embeddings. Transactions of the Association for Computational Linguistics **3** (2015) 211–225

5. Muggleton, S., de Raedt, L.: Inductive logic programming: Theory and methods. The Journal of Logic Programming **19** (1994) 629 – 679
6. Landwehr, N., Passerini, A., Raedt, L.D., Frasconi, P.: kfoil: Learning simple relational kernels. In: AAAI, AAAI Press (2006) 389–394
7. Harnad, S.: The symbol grounding problem. Physica D: Nonlinear Phenomena **42**(1-3) (1990) 335–346
8. Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., Sedivy, J.: Integration of visual and linguistic information during spoken language comprehension. Science **268** (1995) 1632–1634
9. Fillmore, C.J.: Frames and the semantics of understanding. Quaderni di Semantica **6**(2) (1985) 222–254
10. Nüchter, A., Hertzberg, J.: Towards semantic maps for mobile robots. Robot. Auton. Syst. **56**(11) (2008) 915–926
11. Diosi, A., Taylor, G.R., Kleeman, L.: Interactive SLAM using laser and advanced sonar. In: Proceedings of the 2005 IEEE International Conference on Robotics and Automation, ICRA 2005, April 18-22, 2005, Barcelona, Spain. (2005) 1103–1108
12. Bastianelli, E., Bloisi, D.D., Capobianco, R., Cossu, F., Gemignani, G., Iocchi, L., Nardi, D.: On-line semantic mapping. In: Advanced Robotics (ICAR), 2013 16th International Conference on. (Nov 2013) 1–6
13. Chen, D.L., Mooney, R.J.: Learning to interpret natural language navigation instructions from observations. In: Proceedings of the 25th AAAI Conference on AI. (2011) 859–865
14. Matuszek, C., Herbst, E., Zettlemoyer, L.S., Fox, D.: Learning to parse natural language commands to a robot control system. In Desai, J.P., Dudek, G., Khatib, O., Kumar, V., eds.: ISER. Volume 88 of Springer Tracts in Advanced Robotics., Springer (2012) 403–415
15. Bastianelli, E., Castellucci, G., Croce, D., Basili, R., Nardi, D.: Effective and robust natural language understanding for human-robot interaction. In: Proceedings of ECAI 2014, IOS Press (2014)
16. Bastianelli, E., Croce, D., Vanzo, A., Basili, R., Nardi, D.: A discriminative approach to grounded spoken language understanding in interactive robotics. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York. (2016)
17. Altun, Y., Tsochantaridis, I., Hofmann, T.: Hidden Markov support vector machines. In: Proc. of ICML. (2003)
18. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, New York, NY, USA (2004)
19. Collins, M., Duffy, N.: Convolution kernels for natural language. In: Proceedings of Neural Information Processing Systems (NIPS'2001). (2001) 625–632
20. Croce, D., Moschitti, A., Basili, R.: Structured lexical similarity via convolution kernels on dependency trees. In: Proceedings of EMNLP. (2011)
21. Croce, D., Moschitti, A., Basili, R., Palmer, M.: Verb classification using distributional similarity in syntactic and semantic structures. In: 50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference. (2012) 263–272