Crossmodal language grounding, learning, and teaching

Stefan Heinrich, Cornelius Weber, Stefan Wermter
Department of Informatics, Knowledge Technology Group
University of Hamburg
Vogt-Koelln Str. 30, 22527 Hamburg, Germany
{heinrich,weber,wermter}@informatik.uni-hamburg.de

Ruobing Xie, Yankai Lin, Zhiyuan Liu
Department of Computer Science and Technology, Natural Language Processing Lab
Tsinghua University
Room 4-506, FIT Building, Beijing 100084, China
{liuzy}@tsinghua.edu.cn

Abstract

The human brain as one of the most complex dynamic systems enables us to communicate and externalise information by natural language. Despite extensive research, human-like communication with interactive robots is not yet possible, because we have not yet fully understood the mechanistic characteristics of the crossmodal binding between language, actions, and visual sensation that enable humans to acquire and use natural language. In this position paper we present vision- and action-embodied language learning research as part of a project investigating multi-modal learning. Our research endeavour includes to develop a) a cortical neural-network model that learns to ground language into crossmodal embodied perception and b) a knowledge-based teaching framework to bootstrap and scale-up the language acquisition to a level of language development in children of age up to two years. We embed this approach of internally grounding embodied experience and externally teaching abstract experience into the developmental robotics paradigm, by means of developing and employing a neurorobot that is capable of multisensory perception and interaction. The proposed research contributes to designing neuroscientific experiments on discovering crossmodal integration particularly in language processing and to constructing future robotic companions capable of natural communication.

1 Introduction

While research in natural language processing has advanced in parsing and classifying large amounts of text, human-computer communication is still a major challenge: speech recognition is still limited to good signal-to-noise conditions or well adapted models; dialogue systems depend on a well-defined context; and interactive robots that match human communication performance are not yet available. One important reason is the fact that the crossmodal binding between language, actions, and visual events is not yet fully understood and realised in technical systems for the interaction with humans. Imaging techniques such as fMRI have provided a better understanding about which areas in the cortex are involved in natural language processing, and that these areas include somatosensory regions. Language studies have shown that there is a tight involvement of crossmodal sensation and action in speech processing and production as well as in language comprehension. Thus there is

increasing evidence that human language is embodied, which means that it is grounded in most if not all sensory and sensorimotor modalities, and that the human brain architecture favours the acquisition of language by means of crossmodal integration. However, while such language studies shed light on *what* information is processed *where*, they do not address *how* such areas function and compute.

In this position paper we present the vision-and action-embodied language learning research as part of the *Crossmodal Learning* (CML) project. In this project we aim to develop an integrated neural-network and knowledge-based model that processes auditory, visual and proprioceptive information and thus learns language by grounding speech in embodied perception. We develop the neural model, the neurobotic technology to embed it into a real child-like learning environment as well as the teaching framework, which substitutes a teaching caregiver. Such an endeavour will help to understand how information processing and learning takes place spatio-temporally in a cortical model with shared crossmodal representations. Overall, the model and teaching framework will provide a better understanding of how language is learned by grounding speech in embodied perception, how an instruction is interpreted to conduct actions in a real-word scene, and how learning can scale up to a human-like level of language competence.

2 A novel model for embodied language acquisition

Recent findings in neuroscience revealed evidence for embodied language processing in the brain. Specifically, Borghi et al. claimed that the sensorimotor system is involved during perception, action and language comprehension (3). In their review and meta-analysis they inferred that actions as well as words and sentences, which are referring to actions, are firstly encoded in terms of the overall goal (the overall concept) and then of the relevant effectors. In addition, Pulvermüller et al. suggested that for specific combinations of lexical and semantic information a combination of cortical areas, including auditory, motor, or olfactory cortices, can act as binding sites (13).

Latest cortical models in developmental robotics opened up the opportunity to observe language processing on humanoid neurorobots (14; 5; 15), but could not yet achieve a plausibility and complexity with respect to these findings (6). Thus we cannot study the emergence of language and to learn how internal representations for the meaning of utterances develop, and how the temporal flow of information across the modalities is shaped (16). The goal of our research is to address these issues and develop a novel model based on crossmodal and multiple timescale characteristics.

2.1 How can language emerge in a computational model?

The central objectives in developing and studying a neurocognitively plausible model are to understand how language emerges and how the crossmodal integration bootstraps cognition. One specific research question is how temporal dynamic visual and proprioceptive perception contributes to the learning of the meaning of language production. The first hypothesis is that reverberating neural activities act as a memory in extracting a concept for the respective modality and joint modalities.

Secondly, we are keen to examine how the learning of meaning for holistic morphemes (or words) as well as for holo-phrases up to vocabulary-rich phrases scales up in an integrated architecture. The second hypothesis is that a certain degree of embodied grounding is necessary to learn a larger but hierarchically interdependent set of words and phrases. This means that a language learner requires a teaching environment that facilitates the learning of large and rich amounts of examples, including descriptions of possibly grounded or abstract interactions of a child-like learner with its environment (we will elaborate this in Sec. 3).

Validation of these concepts will require to embed the development of such a model as close as possible into a child-like learning environment. Constraints and assumptions can be clarified and limited by employing an embodied robotic implementation.

2.2 Crossmodal grounding in a cortical model

In previous models it was examined how the structure of the human cortex supports the integration of crossmodal input and how these integrative connection patterns result in high-level cognitive capabilities such as language understanding and production, while the development of specific connectivity is based on self-organization from the input data (8; 17). They showed that such a model

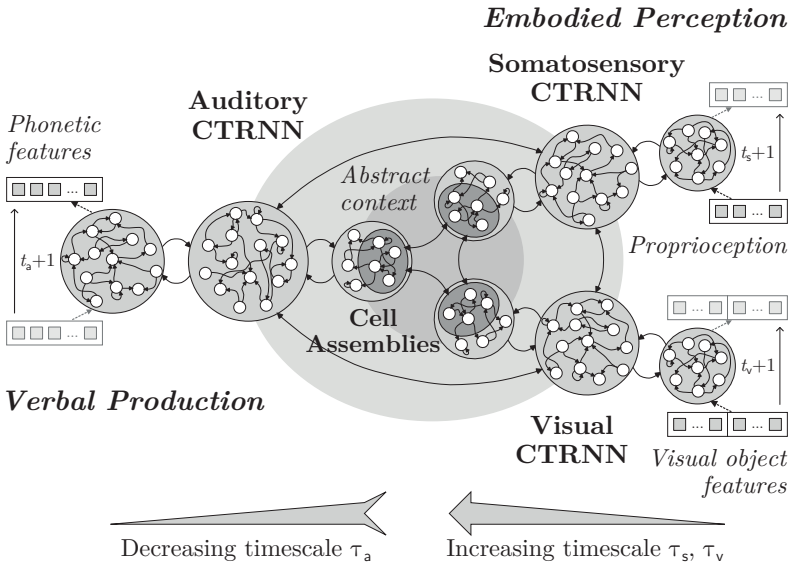

Figure 1: Architecture of the crossmodal model for grounding natural language, consisting of CTRNNs with multiple timescales for auditory production as well as for somatosensory and visual perception, and CAs for representing and processing the primitive concepts as well as abstract concepts. A sequence of phonemes (utterance) is produced over time, based on sequences of embodied crossmodal perception.

can learn sentences by extracting words and the underlying concepts from a verbal utterance and at the same time extract the concept from a body action both from the somatosensory as well as from the visual modalities (8; 9). The abstract concepts for somatosensory and visual perception are quite ambiguous on their own, but could be disambiguated in shared latent representations for the auditory production. To some extent, the model can generalise the verbal description to novel scenes where objects with different shape and colour properties can be recognised. In this paper we propose to develop the model further to allow for an up-scaling to reflect the language competence of up to two year-old children (6).

Our novel model is visualised in Fig. 1. As underlying characteristics, the model is based on a number of key principles:

a The architecture consists of several neurocognitively plausible *Continuous Time Recurrent Neural Networks* (CTRNNs).

b The layers are designed with varying leakage characteristics, thus operating on multiple timescales (19). This is inspired by the findings that a distinct increase in timescale is inherent along the caudal-rostral axis in the frontal lobe (1), indicating that layers of neurons in *higher* level areas process on slow dynamics and high abstraction, whereas neurons in sensory or motor areas process on fast dynamics.

c The layers are on the conceptual level interconnected in *Cell Assemblies* (CAs), thus exchanging and merging the respective auditory, somatosensory, and visual abstract representations (4).

d Shortcut connections are included between intermediate layers in the respective modalities. Although in the brain the motor areas for speech output (controlling muscles for the lips, tongue and the larynx) and somatosensory areas as well as areas involved in visual perception may have different degrees of interconnectivity, the fusion of the information that these areas process happens in higher regions (13). In language processing this indicates that higher level concepts can form by the activation of large and highly distributed CAs of neurons that are strongly and reciprocally connected (11). Other CAs can take on the role of mediators between those concept-CAs and smaller CAs, which represent specific semantics like morphemes and lemmas.

3 A novel knowledgeable language teacher

A growing infant is exposed to many contextually varying discourses to pick up word-segmentation, word-object binding, and generalisation (6). Therefore, the training of a plausible model requires data that stems from some real embodied interactions and is available in large quantities.

To instruct the neural model for embodied language acquisition, we thus propose a novel knowledgeable language teaching system. This artificial teacher is considered to substitute a knowledgeable caregiver that instructs the crossmodal neural model embedded in a learning agent.

3.1 Central goal and functionality

The knowledge-based machine teaching system aims to automatically generate instances to facilitate language learning of the humanoid neurorrobot learner (we will elaborate this in Sec. 4) and must have access to crossmodal observations as well. Since language learning in humans is complex, and requires a child to experience several months of exposure to language, to multisensory context, and to active teaching by a caregiver to understand utterances, it would be extremely time-consuming and impractical for people to act as a language teacher to a robotic learner.

To simulate the real-world teaching scenario between parents and children, the teaching system will automatically provide a large number of crossmodal experiences, containing both possible interactions with objects and their verbal descriptions for neural model learning. In this teaching system, crossmodal information such as images, videos and describing texts are combined for a better overall learning from different aspects. Knowledge graphs are also considered as high-quality structured information for a deeper understanding of the attributes and relationships between objects. This scenario-focused knowledge accounts for core concepts (e.g. objects) and their attributes (e.g. actions afforded by the objects) that may be embodied for the robotic agent. Moreover, the teaching system is supposed to be able to efficiently repeat training sessions with fixed or varying conditions, and also help to avoid the experimenter's bias that is caused by human teachers.

3.2 Methodology

We will implement our crossmodal knowledge-based machine teaching system in a way that the teacher can observe the scene in which the learner interacts, and can describe the scene with a verbal description. The description can be both specific or constructive. Moreover, we will provide crossmodal knowledge in the form of both, visual actions (not necessarily containing embodied somatosensory cues) and verbal descriptions. More specifically, the construction of the teaching system contains four steps:

1. Collect crossmodal information,
2. Develop joint crossmodal knowledge representations,
3. Combine embodied experience with abstract experience of the knowledge-based system,
4. Provide efficient test cases for evaluation.

For the first step, we collect knowledge from a) video streams from the robotic learner interacting with the objects together with a human description of physical respective scene, and b) existing large-scale knowledge bases such as Freebase. Secondly, a joint crossmodal knowledge representation model will be developed to better understand potential attributes and relations between objects. The representation of images and video streams could be inspired by the huge success of convolutional neural networks for a variety of tasks like image classification (10), while the representation of knowledge could be learned under the state-of-the-art framework of translation-based methods (12). Moreover, we will also exploit several techniques (such as efficient object detection and entity recognition) to enrich video streams with text knowledge, to simulate further experience.

As the third step, we combine embodied interactions with abstract knowledge from the knowledge-based teaching system into the teaching process. The robotic learner must be able to bridge the gap from words and concepts learned from embodied interaction to additional knowledge learned from the teaching system. Thus, related examples will be provided that the learner has only experienced visually or somatosensorily, or may have never experienced at all. For example, the learner may experience a red cup and learn its attributes (both how a cup looks according to its shape and colour

and how it feels by touching). In addition, the teaching system will enable the robotic learner to extensively learn about various types of cups with different shapes and colours (or even teacups and bowls) from images or videos.

Finally, to evaluate whether the robotic learner can develop appropriate understandings and motor actions, the knowledge-based teacher should also provide efficient test cases. Since we simulate the real-world language teaching, the test cases will be embedded in a crossmodal scenario.

3.3 Significance and feasibility

Knowledge is essential and plays a significant role in the teaching system. First, we suppose that structured knowledge reflects the cognition to the real world. Cognition could compose discrete learned information to a structured knowledge graph, linking objects to each other or to their attributes, which helps language learning. Second, the joint knowledge representation enables association to similar objects (e.g. apple and orange) even if are never experienced by the learner. Knowledge inference is also available via the crossmodal representations, so that we can better understand more complicated instructions. Third, structured knowledge provides information for disambiguation, which is an essential challenge in language learning and understanding. It is natural to construct structured knowledge simultaneously while learning language.

Previous work on knowledge representation builds the foundation of our knowledge-based teaching system. TransE projects both, entities and relations, into the same continuous low-dimensional vector space, interpreting relations as translating operations between head and tail entities (2). This helps us to build relationships between entities, which improves the ability of association in language learning. As for crossmodal learning in knowledge representation, DKRL proposes a novel description-based representation for each entity (18), which combines textual information in entity descriptions with structured information in knowledge graphs. Our method of a crossmodal knowledge representation could be inspired by this framework. These recent approaches will form a basis to explore scenario-focused crossmodal knowledge extraction to enable a robot to learn languages.

4 Technologies and experimental design

Embodied language learning of the neurocognitive model will be studied on an interactive humanoid robot platform that is instructed by the knowledgeable teacher. As a first important characteristic, the learner must feature multimodal sensation capabilities to emulate the rich perception of a learning infant. For our particular scenario, the learner is supposed to interact with different objects in its field of view and must capture continuous stereo-visual and somatosensory perception – specifically proprioceptive information for an arm, but also tactile information for the hand – as well as perceive verbal descriptions from the knowledgeable teacher (see Fig. 2). The second important characteristic is the knowledgeable teacher that can observe and interpret the interactions and also inform about knowledge that is partially related to the interaction or even disconnected at all. The platform is overall related to the developmental robotics approach, in terms of capturing developmental principles and mechanisms observed in the natural cognitive systems of children (6), but is designed bottom-up from the neurocognitive model (8).

4.1 Evaluation and analysis methodology

To analyse the model's characteristics, we are interested to identify parameter settings for the best (relative) generalisation capabilities in order to analyse the information patterns that emerges for different parts of the architecture. Inspired from infant learning the evaluation will be embedded in the real world scenario, where the robot is supposed to show ideal performance for novel but related scenes, compared to the learned natural language interactions with the teacher and its environment (6). The analysis will be focused on correlating and quantifying the latent representations that might form in different higher level areas for specific modalities and in the higher level as well as mediating CAs between the modalities. This methodology is related to analysing representations from MEG or fMRI data (7), and can reveal how the architecture self-organises, based on the spatio-temporal dynamics as well as the hierarchical dependencies in the multi-modal training data.

Figure 2: The learner explores interaction with objects, while the knowledgeable teacher describes the respective interaction in natural language.

4.2 Up-scaling and meaningful uncertainty in a learner-caregiver scenario

In order to scale the model, the neurorobotic learner will be exposed to rich permutations of temporal dynamic multimodal perception and desired auditory productions, thus to multi-variance data with a tremendous complexity and seemingly large degree of uncertainty. A central hypothesis for the learning is that a neurocognitively plausible model might be able to self-organise the binding of coherent sounds (or words) with visual entities or certain arm movements (primitive actions), but could heavily benefit from scaffolding: Similar to language learning in infants the learning could be shaped by teaching simple words and holo-phrases first, and then more complex utterances – without altering weights for certain layers in the architecture. In preliminary tests we learned that layers with a medium timescale show a predisposition in capturing holistic words and can be used to generate novel phrases. In addition, such step-by-step learning could also reveal differences and similarities in developed internal representations that emerge from learning.

5 Conclusion

We propose a research approach for understanding the temporal dynamics and mechanism underlying language processing and acquisition in the human brain. The development and study of our cortical neural model, integrated in a knowledgeable teaching framework, can provide theoretical and experimental evidence for the general hypothesis that a neural architecture that maximizes crossmodal as opposed to unimodal representations requires fewer resources and learning cycles, and can achieve better generalisation to unknown circumstances but also greater robustness in memorising and accessing its stored representations. With such an outcome we can design novel neuroscientific experiments on discovering crossmodal integration particularly in language processing and construct future robotic companions that provide better communication capabilities.

Acknowledgments

The authors gratefully acknowledge support from the German Research Foundation (DFG) and the National Science Foundation of China (NSFC) under project Crossmodal Learning, TRR-169.

References

[1] D. Badre, A. S. Kayser, and M. D'Esposito. Frontal cortex and the discovery of abstract action rules. *Neuron*, 66(2):315–326, 2010.

[2] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko. Translating embeddings for modeling multi-relational data. In *Proc. NIPS 2013*, pages 2787–2795, 2013.

[3] A. M. Borghi, C. Gianelli, and C. Scorolli. Sentence comprehension: effectors and goals, self and others. An overview of experiments and implications for robotics. *Frontiers in Neurorobotics*, 4(3):8 p., 2010.

[4] V. Braitenberg and A. Schüz. *Cortex: Statistics and geometry of neuronal connectivity*. Springer-Verlag Berlin Heidelberg, 1998.

[5] A. Cangelosi. Grounding language in action and perception: From cognitive agents to humanoid robots. *Physics of Life Reviews*, 7(2):139–151, 2010.

[6] A. Cangelosi and M. Schlesinger. *Developmental robotics: From babies to robots*. The MIT Press, 2015.

[7] R. M. Cichy, A. Khosla, D. Pantazis, T. Antonio, and A. Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, 6:13 p., 2016.

[8] S. Heinrich and S. Wermter. Interactive language understanding with multiple timescale recurrent neural networks. In *Proc. 24th ICANN*, volume 8681 of *LNCS*, pages 193–200, 2014.

[9] S. Heinrich and S. Wermter. Interactive natural language acquisition in a multi-modal recurrent neural architecture. *Connection Science*, under review, 2017.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. NIPS 2012*, pages 1097–1105, 2012.

[11] W. J. M. Levelt. Spoken word production: A theory of lexical access. *Proceedings of the National Academy of Sciences of the United States of America*, 98(23):13464–13471, 2001.

[12] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu. Learning entity and relation embeddings for knowledge graph completion. In *Proc. AAAI-15*, pages 2181–2187, 2015.

[13] F. Pulvermüller, M. Garagnani, and T. Wennekers. Thinking in circuits: toward neurobiological explanation in cognitive neuroscience. *Biological Cybernetics*, 108(5):573–593, 2014.

[14] D. K. Roy. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1–2):170–205, 2005.

[15] L. Steels and M. Hild. *Language Grounding in Robots*. Springer Science+Business Media LLC New York, 2012.

[16] J. Tani. Self-organization and compositionality in cognitive brains: A neurorobotics study. *Proceedings of the IEEE*, 102(4):586–605, 2014.

[17] M. Vavrečka and I. Farkaš. A multimodal connectionist architecture for unsupervised grounding of spatial language. *Cognitive Computation*, 6(1):101–112, 2014.

[18] R. Xie, Z. Liu, J. Jia, H. Luan, and M. Sun. Representation learning of knowledge graphs with entity descriptions. In *Proc. AAAI-16*, pages 2659–2665, 2016.

[19] Y. Yamashita and J. Tani. Emergence of functional hierarchy in a multiple timescale neural network model: A humanoid robot experiment. *PLOS Computational Biology*, 4(11):e1000220, 2008.