Interactive Learning and Cross-Modal Binding - A Combined Approach*

Henrik Jacobsson[1], Nick Hawes[2], Danijel Skočaj[3], Geert-Jan M. Kruijff[1]
[1] Language Technology Lab, DFKI GmbH, Germany, [2] School of Computer Science, University of Birmingham, UK
[3] University of Ljubljana, Slovenia,

Introduction

To function properly in the world, a cognitive system should possess the ability to learn and adapt in a continuous, open-ended, life-long fashion. This learning is inherently cross-modal; the system should use all of its percepts and capabilities to sense and understand the environment, and update the current knowledge accordingly. For the life-long learning to be effective, it is also important to be able to incorporate knowledge from other knowledgable cognitive systems through *interactive learning*. For this to be "socially acceptable", it is important to support a wide variety of tutoring channels. For example, to treat the tutor only as a source for linguistic labels is not a natural way of communication and is thus not very effective from the human's point of view. For an excellent and deep account of proper design considerations for socially interactive learning systems see (Thomaz, 2006)

A prerequisite for interactive learning is the successful interpretation of the meaning of references used in dialogue with a human. The robot must therefore be able to form associations between information in different modalities, e.g., between linguistic references and visual input (Roy, 2005). Forming these associations is a process we refer to as *cross-modal binding*. We are developing a multifaceted approach to binding, and in this extended abstract we address the virtue of the symbiosis of binding and interactive learning.

Cross-Modal Binding

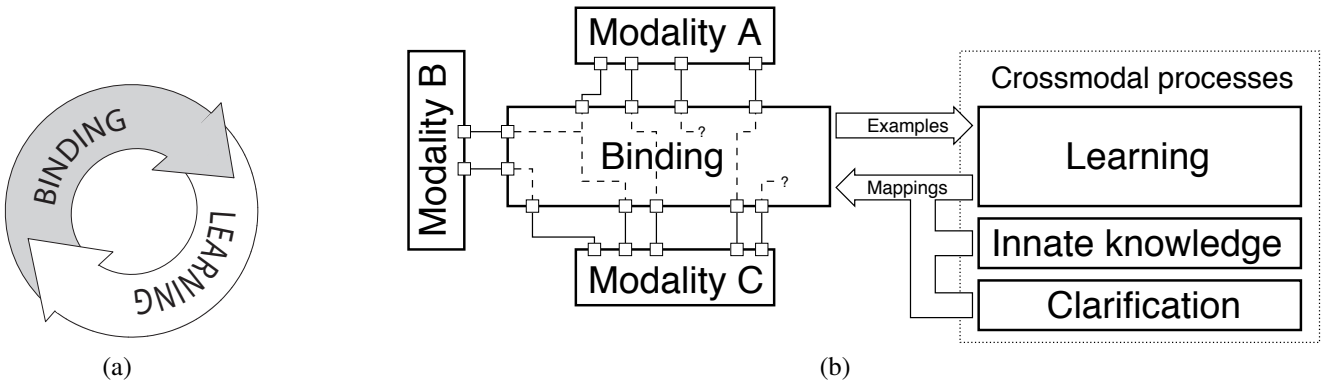
We treat the binding of linguistic and visual content as an instance of a broader cross-modal binding problem: to enable a broad and open-ended set of modalities to contribute towards a common representation of abstract concepts, objects, and actions, and N-ary relations between them. For example, for a robot to successfully determine the correct response to the "give me the blue mug that's to the right of the plate" it must be able to correctly interpret the references to the objects, the action, and the spatial relationship.

*This work was supported by the EU FP6 IST Cognitive Systems Integrated Project "CoSy" FP6-004250-IP.

Typical robotic systems are composed of specialised subsystems, e.g. vision, manipulation, dialogue, reasoning etc. For N subsystems there are $N(N-1)/2$ potential interfaces between them. Building associations in this manner can quickly become expensive to manage both at design- and run-time. To avoid this, we employ a two-level approach to binding. The bottom level corresponds to subsystem specific representations. The second level represents objects, actions and relations by bundling together sets of *features* abstracted from the first level representations. These "bundles" represent a subsystem's best hypotheses about the objects, actions and relations in its modality. To build a common representation from all its subsystems, a number of *binding processes* then operate on this more abstract level of information. This is illustrated in Figure 1. Further information is available in previous work (Hawes et al., 2007). The focus of this abstract is that the information used to associate features across modalities may be learned, and that this two-level system naturally supports such learning.

Cross-Modal Learning

When the binding processes establish associations between bundles of abstracted features, these associations implicitly link features from these modalities. Some of these links will represent known cross-modal mappings between features, but others may represent valid mappings that the system does not know about. For example, in the utterance "give me the blue mug that's to the right of the plate" visual colour features (blue pixels) may be implicitly linked to linguistic colour features ("blue") via an association formed from a type description ("the mug"). When the binding of the object descriptions succeed, the binder can generate novel training examples for a learning module. In the case above, the binder would generate the training examples for updating the representations of "blue", "the mug", "to the right of", and "the plate". In this way, the system can increase its current knowledge without being explicitly instructed, and without training examples being provided separately. An idealised learner would try to use all the inferred information and data from all modalities to *co-train* (cf. Levin et al.,



(a) (b)

Figure 1: Cross-modal binding and incremental learning are tightly integrated (a) and feed each other information regarding mappings and learning examples respectively. They are also integrated in the context of the modalities themselves as well as other cross-modal processes that can assist the binder prior to, and in parallel with, learning (b). Clarification is one way of explicitly establishing mappings, e.g. through dialogue. Over time, the dual interaction between these subsystems results in increasingly informative learning examples and mapping suggestions.

2003) its representations in other modalities as well.

Any learning method using binding processes for training will be thus fed by a stream of examples of cross-modal associations. The open-ended nature of this input makes it important that any learning systems used are *incremental*; the learning process should continue to improve the learned models from the incoming examples. The learned models should also be capable of determining their own level of confidence so that they contribute to the binder only when the representations are sufficiently stable and confident.

Since the learners can potentially know what kind of examples will help them train, they may also trigger explicit clarification behaviour. This will allow more focused clarification behaviour than the binder could trigger on its own (for ambiguous bindings etc.). For example, if a learned colour classifier is uncertain about labelling a visual object as "orange" or "red", it could trigger a clarification behaviour to distinguish between *precisely* these labels. Or it could even issue a command to the manipulation module to turn the object in order to provide additional visual information to resolve ambiguity.

Taken together, the learning, binding and clarification behaviours of the system form a strong basis for a wide range of tutor-robot interactions. There is potential for mixed initiative dialogue since the agent will autonomously ask for information. Despite this, the learning is essentially autonomous and can passively listen for examples in case the agent is engaged in other behaviours. Additionally, by using abstracted representations of subsystem-specific representations, many of the mechanisms for reasoning and learning can be reused across very different domains. For example, general-purpose clarification planning mechanisms can be employed across different modalities.

Conclusion

In our approach, the binding processes and learning processes thus form a symbiosis, where they both benefit from the information they feed each other. The learners benefit from helping the binder to make cross-modal associations since it will then be able to make more (and potentially better) bindings, potentially resulting in more training data for the learners. By supporting the learning of cross-modal associations, e.g. between colour labels and visual colour representations, the binder may implicitly be able to associate information from other feature spaces, e.g. about spatial relations, via these learnt assoications. Moreover, an intelligent robot employing this approach also benefits from the symbiosis; its primary source of cross-modal mapping information (information required to support linguistic interaction) can be shifted from costly, deliberate tutoring to incrementally learned associations.

References

Hawes, N., Sloman, A., Wyatt, J., Zillich, M., Jacobsson, H., Kruijff, G., Brenner, M., Berginc, G., and Skočaj, D. (2007). Towards an integrated robot with multiple cognitive functions. In *Proc. AAAI '07*.

Levin, A., Viola, P. A., and Freund, Y. (2003). Unsupervised improvement of visual detectors using co-training. In *ICCV*, pages 626–633, Nice, France.

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

Thomaz, A. L. (2006). *Socially Guided Machine Learning*. PhD thesis, Massachusetts Institute of Technology.