In *Proceedings of the 2006 AAAI Fellows Symposium*. Boston, MA, July 2006.

# Learning Language from Perceptual Context: A Challenge Problem for AI

**Raymond J. Mooney**
Department of Computer Sciences
University of Texas at Austin
1 University Station C0500
Austin, TX 78712-0233

**Abstract**

We present the problem of learning to understand natural language from examples of utterances paired only with their relevant real-world context as an important challenge problem for AI. Machine learning has been adopted as the most effective way of developing natural-language processing systems; however, currently, complex annotated corpora are required for training. By learning language from perceptual context, the need for laborious annotation is removed and the system's resulting understanding is grounded in its perceptual experience.

## Introduction

Understanding and communicating in natural language is one of the defining problems of AI, as indicated by the observation that it is the core task underlying Turing's famous test. Over the past ten to fifteen years, there has been a revolution in the way natural-language processing (NLP) systems are developed. In early NLP systems, the phonetic, syntactic, semantic, and pragmatic knowledge required for language understanding was developed manually by computational linguists. However, assembling and encoding such knowledge was found to be arduous and error-prone. These days, almost all effective NLP systems rely extensively on the use of machine learning techniques to automatically acquire the knowledge they require from linguistic training data. Due to the increased robustness and coverage engendered by learning approaches, the so-called "empirical," "statistical," or "corpus-based" approach has almost completely supplanted manual knowledge engineering in NLP.

However, current learning methods for NLP require annotating large text corpora with supervisory information such as part-of-speech tags, syntactic parse trees, semantic role labels, word senses, etc.. Building corpora such as the Penn tree bank (Marcus, Santorini, & Marcinkiewicz 1993) is an expensive, arduous task. As one moves towards deeper semantic analysis, the annotation task becomes increasingly more difficult and complex. Our prior and on-going research has developed techniques for learning semantic parsers that map natural-language sentences into a formal meaning representation such as first-order predicate logic (Zelle & Mooney 1996; Ge & Mooney 2005; Wong & Mooney 2006; Kate & Mooney 2006). However, in order to make the annotation and learning task tractable, we have restricted our work to specific applications, such as answering natural-language queries to a specific database, or interpreting Robocup soccer coaching instructions.

Ideally, an AI system would be able to learn language like a human child, by being exposed to utterances in a rich perceptual context. By inferring the meaning of a sentence from the context in which it was uttered, a sentence-meaning pair could be automatically constructed. Methods for inducing semantic parsers from sentences annotated with meaning representations could then be applied to the resulting data. Although in general it is not possible to infer a unique meaning for a sentence from context, in the vast majority of cases, the context greatly restricts its range of possible meanings.

I believe the time is ripe to make a serious attempt at tackling the problem of learning language from natural context. The individual fields of computational linguistics, machine learning, knowledge representation, computer vision, and robotics have reached a level of maturity that I believe a serious attempt at attacking this problem is now viable. However, pursuing such an agenda does require collaboration between these currently very distinct and separated areas of AI. Consequently, this challenge problem is a perfect venue for attempting to re-integrate these and other areas of AI that, unfortunately, have grown further and further apart.

## Relevant Existing Research

The general problem of *symbol grounding*, how the meaning of abstract symbols is grounded in an agent's perceptual environment and experience, has been argued to be a critical issue in developing truly intelligent artificial systems (Harnad 2004). Clearly, a deep understanding of most natural language requires capturing the connection between the abstract concepts underlying words and phrases and their embodiment in the physical world. Even most abstract concepts are based on metaphorical references to more concrete physical concepts (Lakoff & Johnson 1980).

There has been some recent work on inferring a grounded meaning of individual words or short referring expressions from visual perceptual context (Roy 2002; Bailey *et al.* 1997; Barnard *et al.* 2003; Yu & Ballard 2004). However, the syntactic complexity of the natural language used in this work is very restrictive, many of the systems use existing knowledge of the language, and most of them use static images to learn language describing objects and cannot use dynamic video to learn language describing actions. None of this existing work makes use of modern statistical-NLP parsing techniques or learns to build detailed symbolic meaning representations of complete, complex sentences.

## The Way Forward

I believe there are a number of productive avenues for making progress on addressing the proposed challenge problem. Pursuing the problem from multiple directions, addressing the important issues in each of the areas of NLP, machine learning, computer vision, and robotics will be critical, with the eventual goal of integrating the insights and techniques learned from research in these different areas. Below, I briefly describe some of my own plans for tackling the problem from the perspective of semantic-parser acquisition, by learning to map sentences into formal meaning representations by training on examples of language paired with abstracted, symbolic descriptions of its perceptual context.

Part of the difficulty of doing research in learning language from perceptual context is that it requires knowledge and skills in both NLP and computer vision, and possibly also speech recognition and robotics. The complexity of building a complete system is beyond the expertise of any individual or small group. Since the researchers who have done existing work in the area are not NLP experts, the complexity of the language involved has been quite modest. In order to make progress on the problem from an NLP perspective, I believe it will be productive to study the problem in simulated environments that retain many of the important properties of a dynamic world with multiple agents and actions while temporarily avoiding many of the complexities of vision processing. Specifically, I have made initial plans to use the Robocup simulator (Chen *et al.* 2003) which provides a fairly detailed physical simulation for robot soccer. Several groups have constructed Robocup "commentator" systems (André *et al.* 2000) that provide a natural-language transcript of the simulated game, such as that produced by a sports announcer. The goal of our initial project is to construct a system that learns to semantically interpret language in this domain by observing such an on-going language description of the activity on the field paired with the corresponding dynamic simulator state. By exploiting existing techniques for abstracting a symbolic description of the activity on the field from the detailed state of the physical simulator (André *et al.* 2000), we can obtain a pairing of natural language with a detailed symbolic description of the perceptual context in which it was uttered. This will allow for in-depth, controlled study of interesting problems in learning language from dynamic perceptual context while avoiding the limitations and complexity of existing vision and robotic systems. Language learning in the context of simulated interactive video-game environments (Fleischman & Roy 2005) provides similar advantages.

In addition to the approaches taken in the existing research referenced above, other promising initial approaches might include learning from the simple language and corresponding pictures in children's books, learning individual word meanings from symbolic descriptions of context (Siskind 1996), and learning to abstract symbolic descriptions of objects and actions from video input (Fern, Givan, & Siskind 2002).

## References

André, E.; Binsted, K.; Tanaka-Ishii, K.; Luke, S.; Herzog, G.; and Rist, T. 2000. Three RoboCup simulation league commentator systems. *AI Magazine* 21(1):57–66.

Bailey, D.; Feldman, J.; Narayanan, S.; and Lakoff, G. 1997. Modeling embodied lexical development. In *Proc. of 19th Annual Conf. of the Cognitive Science Society*.

Barnard, K.; Duygulu, P.; Forsyth, D.; de Freitas, N.; Blei, D. M.; and Jordan, M. I. 2003. Matching words and pictures. *Journal of Machine Learning Research* 3:1107–1135.

Chen, M.; Foroughi, E.; Heintz, F.; Kapetanakis, S.; Kostiadis, K.; Kummeneje, J.; Noda, I.; Obst, O.; Riley, P.; Steffens, T.; Wang, Y.; and Yin, X. 2003. Users manual: RoboCup soccer server manual for soccer server version 7.07 and later. Available at `http://sourceforge.net/projects/sserver/`.

Fern, A.; Givan, R.; and Siskind, J. M. 2002. Specific-to-general learning for temporal events with application to learning event definitions from video. *Journal of Artificial Intelligence Research* 17:379–449.

Fleischman, M., and Roy, D. 2005. Intentional context in situated natural language learning. In *Proc. of 9th Conf. on Computational Natural Language Learning (CoNLL-2005)*, 104–111.

Ge, R., and Mooney, R. J. 2005. A statistical semantic parser that integrates syntax and semantics. In *Proc. of 9th Conf. on Computational Natural Language Learning (CoNLL-2005)*, 9–16.

Harnad, S. 2004. The symbol grounding problem. *Physica D* 42:335–346.

Kate, R., and Mooney, R. J. 2006. Using string-kernels for learning semantic parsers. In *Joint Conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING-ACL-06)*.

Lakoff, G., and Johnson, M. 1980. *Metaphors We Live By*. Chicago: University Of Chicago Press.

Marcus, M.; Santorini, B.; and Marcinkiewicz, M. A. 1993. Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics* 19(2):313–330.

Roy, D. 2002. Learning visually grounded words and syntax for a scene description task. *Computer Speech and Language* 16(3):353–385.

Siskind, J. M. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition* 61(1):39–91.

Wong, Y., and Mooney, R. J. 2006. Learning for semantic parsing with statistical machine translation. In *Proceedings of the 2006 Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT/NAACL-06)*.

Yu, C., and Ballard, D. H. 2004. On the integration of grounding language and learning objects. In *Proc. of 19th Natl. Conf. on Artificial Intelligence (AAAI-2004)*, 488–493.

Zelle, J. M., and Mooney, R. J. 1996. Learning to parse database queries using inductive logic programming. In *Proc. of 13th Natl. Conf. on Artificial Intelligence (AAAI-96)*, 1050–1055.