Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008)

Learning to Connect Language and Perception

Raymond J. Mooney
Department of Computer Sciences
University of Texas at Austin
1 University Station C0500
Austin, TX 78712-0233
mooney@cs.utexas.edu

Abstract

To truly understand language, an intelligent system must be able to connect words, phrases, and sentences to its perception of objects and events in the world. Current natural language processing and computer vision systems make extensive use of machine learning to acquire the probabilistic knowledge needed to comprehend linguistic and visual input. However, to date, there has been relatively little work on learning the relationships between the two modalities. In this talk, I will review some of the existing work on learning to connect language and perception, discuss important directions for future research in this area, and argue that the time is now ripe to make a concerted effort to address this important, integrative AI problem.

Introduction

Currently, the most effective methods for natural-language processing (NLP) make extensive use of machine learning to acquire the probabilistic knowledge needed to recognize words in speech, parse sentences, disambiguate words, and translate between languages (Manning & Schütze 1999). However, current learning methods for NLP require annotating large corpora with supervisory information such as textual transcriptions of speech, part-of-speech tags, syntactic parse trees, semantic role labels, word senses, and parallel bilingual text. Building such corpora is an expensive, arduous task. As one moves towards deeper semantic analysis, the annotation task becomes increasingly more difficult and complex. In my prior research, I have developed techniques for learning semantic parsers that map natural-language sentences into a formal meaning representation such as first-order logic (Zelle & Mooney 1996; Ge & Mooney 2005; Kate & Mooney 2006; Wong & Mooney 2007b). The training data for these systems consists of a corpus of sentences each paired with a formal representation of its meaning. However, to make the annotation tractable, we restricted our work to specific applications, such as answering natural-language database queries or interpreting Robocup (www.robocup.org) coaching instructions.

Also, traditional semantic representations in NLP consist of connections between words or formulae with arbitrary conceptual symbols. A true understanding of natural language requires capturing the connection between linguistic concepts and the world as experienced through perception, i.e. *symbol grounding* (Harnad 1990). Even most abstract concepts are based on metaphorical references to more concrete physical concepts (Lakoff & Johnson 1980).

Ideally, an AI system would be able to learn language like a human child, by being exposed to utterances in a rich perceptual environment. The perceptual context would provide the necessary supervisory information, and learning the connection between language and perception would ground the system's semantic representations in its perception of the world. However, such supervisory information is highly ambiguous and noisy, since the language could be referring to numerous aspects of the environment (William James' "blooming, buzzing confusion"), or may not even refer to anything in the current perceivable context.

The most effective current methods for computer vision also make extensive use of machine learning to acquire the probabilistic knowledge needed to segment images, recognize objects in images, and detect events in videos (Sebe *et al.* 2005). However, these methods also typically require laboriously constructed training sets of human-annotated images and video. Ideally, naturally accompanied linguistic data, e.g. captions for images and video, would be able to effectively substitute for some of this explicit supervision.

I believe the time is ripe for exploring the task of learning the connection between language and perception. The individual fields of computational linguistics, machine learning, computer vision, and robotics have all reached a level of maturity that I believe a serious attempt to address this problem is now viable. However, pursuing such an agenda requires collaboration between these distinct areas of AI. Consequently, this challenge problem is a perfect venue for attempting to re-integrate important subareas of AI that, unfortunately, have grown further and further apart.

First, this paper briefly reviews some of the existing work in the area, next it describes some of my own on-going research on the problem, and finally it concludes with a discussion of important directions for future research.

Review of Prior Research

Although, to date, learning to connect language and perception has not been extensively studied, several systems have

been developed that infer grounded meanings of individual words or descriptive phrases from their visual context. Deb Roy and his research group have built several robots that learn linguistic descriptions of objects (Roy 2005). In particular, he has shown that by detecting correlations between audio and visual modalities in the context of infant-directed speech, that a system could learn to segment speech into discrete words more accurately than using audio input alone (Roy & Pentland 2002). Roy (2002) also developed a system that learns to describe simple geometrical figures in an image based on their size, color, and shape as well as their location relative to other figures in the scene.

Yu and Ballard (2004) have also developed methods that learn to associate nouns with the perception of particular objects. The NLT group at Berkeley has worked on learning linguistic descriptions of "blocks world" scenes (Feldman *et al.* 1996). Barnard *et al.* (2003) developed a method that learns to associate words with image regions or complete pictures by training on captioned images. Barnard and Johnson (2005) present a method that learns to choose the correct sense of an ambiguous word based on an accompanying image, such as picking a meaning for "bank" when it is paired with either a picture of a building or a river. There is also a growing body of recent work in computer vision that uses text captions to help classify and cluster images on the web, e.g. (Bekkerman & Jeon 2007). However the language used in existing work is quite restrictive and the vision, neural-net, and robotics researchers who have conducted this research did not exploit the latest statistical NLP methods for learning complex natural-language grammar.

Our Current Research

As an initial test of grounded language learning, we have recently developed a system that learns to sportscast simulated soccer games by training on sample "play by play" commentaries (Chen & Mooney 2008). We decided to initially study the problem in a simulated environment that retains the properties of a dynamic world with multiple agents and actions while avoiding the complexities of computer vision. Specifically, we use the Robocup simulator[1] which provides a fairly detailed physical simulation of robot soccer. Our system learns to interpret and generate language in the Robocup domain by observing on-going textual commentary on a game together with the evolving state of the simulator. By exploiting existing techniques for abstracting a symbolic description of the activity on the field from the simulator state (André *et al.* 2000), we obtain a pairing of natural language with a symbolic description of the perceptual context in which it was uttered. However, such training data is highly ambiguous because a typical comment co-occurs with multiple events to which it could refer. We enhanced our existing methods for learning semantic parsers and language generators (Kate & Mooney 2007; Wong & Mooney 2007a) in order to learn to understand and produce language from such ambiguous training data. A screenshot of our system with generated commentaries is shown in Figure 1.

[1] `http://sourceforge.net/projects/sserver/`

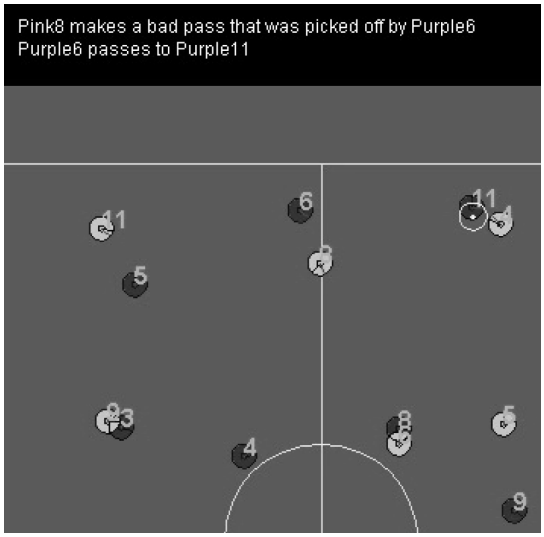

Figure 1: Screenshot of our Robocup Sportscasting System

The system uses ideas from our system KRISPER (Kate & Mooney 2007) to learn from the weakly-supervised, ambiguous training data. KRISP (Kate & Mooney 2006) uses *support-vector machines* (SVMs) with *string kernels* (Lodhi *et al.* 2002) to learn semantic parsers. KRISPER extends KRISP to handle ambiguous training data, in which each natural-language (NL) sentence is annotated only with a *set* of potential meaning representations (MRs), only one of which is correct. It employs an iterative approach analogous to *expectation maximization* (EM) that improves upon the selection of the correct NL–MR pairs in each iteration. In the first iteration, it assumes that all of the MRs paired with a sentence are correct and trains KRISP with the resulting noisy supervision. In subsequent iterations, KRISPER uses the currently trained parser to score each potential NL–MR pair, selects the most likely MR for each sentence, and retrains the parser.

Our sportscasting system also uses ideas from our system WASP (Wong & Mooney 2006; 2007b) in order to learn how to *generate* language as well as to parse it. WASP uses *statistical machine translation* (SMT) techniques to induce a *probabilistic synchronous context-free grammar* (PSCFG) (Wu 1997) to translate NL sentences into logical MRs. The PSCFG is learned using a modification of recent methods in syntax-based SMT (Chiang 2005). Since a PSCFG is symmetric with respect to input/output, the same learned model can also be used to *generate* NL sentences from formal MRs. Thus, WASP learns a PSCFG that supports both semantic parsing and natural language generation. By using the EM-like retraining procedure used in KRISPER, we developed a version of WASP that learns to parse and generate language given only ambiguous training data.

However, a language generator alone is not enough to produce a sportscast. In addition to knowing *how* to say something, one must also know *what* to say. A sportscaster must also choose which events to describe. In NLP, deciding what to say is called *strategic generation*. We also developed a simple approach to learn, from the available ambiguous

training data, the types of events that a sportscaster is most likely to describe.

To train and test our system, we assembled commentated soccer games from the Robocup simulation league (www.robocup.org) by having humans commentate games while watching them on the simulator. We annotated a total of four games, namely, the finals for each year from 2001 to 2004. We then trained our system on three games and tested it on the fourth game, averaging over the four possible test games. We evaluated how well the system learned to parse sentences into MRs and generate language describing game events, as well as its ability to determine which sentences referred to which game events and to decide which game events to comment on. We also presented human evaluators a random mix of human-commentated and machine-commentated game clips and asked them to subjectively rate their quality on a 5-point scale. On average, the machine-commentated clips received a score only 0.7 points below the human ones. Although we did not quite achieve human-level performance, we believe the results clearly demonstrate that our system can effectively perform grounded language learning from a very limited amount of realistic training data. Chen and Mooney (2008) present the detailed results.

Directions for Future Research

There are important problems in each of the areas of NLP, machine learning, and computer vision that need to be addressed to make progress on the proposed challenge, with the eventual goal of integrating the techniques developed. From NLP, effective methods are needed for learning from language paired with semantic representations. This has been one of the foci of my own research over the last decade, and a few other researchers have started to explore the problem (Zettlemoyer & Collins 2005).

From machine learning, methods are needed that can learn from highly ambiguous training data in which sentences are paired with many potential meanings only one (if any) of which is correct. A few initial approaches to handling such *referential uncertainty* or *ambiguous training data* have been proposed (Siskind 1996; Kate & Mooney 2007); however, methods are needed that exploit more clues and constraints, such as those studied in psychological research on child language acquisition (Bloom 2000). Another interesting connection that should be explored is the one between language learning from such ambiguous supervision and *multiple instance learning* (MIL) (Dietterich, Lathrop, & Lozano-Perez 1997; Ray & Craven 2005). In multiple instance learning, weak supervision is provided in the form of "bags" of instances of which at least one is guaranteed to be a positive example of the target concept. If given sentences with sets of possible meaning representations, it constitutes a form of MIL for *learning with structured data* (Bakir *et al.* 2007). In this case, training consists of input strings each paired with a set of possible structured meaning representations of which at exactly one is correct. To my knowledge, there is currently no work on MIL for structured output.

From computer vision, methods are needed for learning to recognize objects and events from visual input. In particular, there has been significant recent progress in identifying intentional human actions in video input (Fern, Givan, & Siskind 2002; Aggarwal & Park 2004); however, significant additional development is required to extract the range of objects and events to which language can refer. Ideally, methods are needed for using naturally accompanying linguistic clues as a substitute for explicit supervisory labels.

Finally, good testbeds and datasets are needed that can serve as shared experimental benchmarks. Simulated environments, like that used in our Robocup sportscasting task, will be useful for studying the NLP and learning issues without first having to solve all of the difficult computer vision problems. In particular, interactive video-games can provide rich simulated environments for grounded language learning (Fleischman & Roy 2005; Gorniak & Roy 2005). Another promising benchmark is learning from the language and corresponding drawings in children's picture books. Images and video with text captions can also provide a rich source of useful data, e.g. (Fleischman & Roy 2007).

In conclusion, I believe that learning to connect language and perception raises a number of interesting challenge problems that requires integrating ideas and methods from machine learning, NLP, computer vision, and robotics. I believe that substantial progress can be made on these important problems in the ensuing years, and I would encourage others to join me in exploring this exciting research direction.

Acknowledgements

I would like to thank my current and former graduate students, John Zelle, Cindi Thompson, Yuk Wah (John) Wong, Rohit Kate, Ruifang Ge, and David Chen for contributing substantially to the work described in this paper. This research was supported by the National Science Foundation through grant IIS–0712097.

References

Aggarwal, J. K., and Park, S. 2004. Human motion: Modeling and recognition of actions and interactions. In *Proceedings of the 2nd International Symposium on 3D Data Processing, Visualization, and Transmission*.

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.

Bakir, G.; Hoffman, T.; Schölkopf, B.; Smola, A. J.; Taskar, B.; and Vishwanathan, S. V. N., eds. 2007. *Predicting Structured Data*. Cambridge, MA: MIT Press.

Barnard, K., and Johnson, M. 2005. Word sense disambiguation with pictures. *Artificial Intelligence* 167:13–30.

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.

Bekkerman, R., and Jeon, J. 2007. Multi-modal clustering for multimedia collections. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*.

Bloom, P. 2000. *How Children Learn the Meanings of Words*. Cambridge, MA: MIT Press.

Chen, D. L., and Mooney, R. J. 2008. Learning to sportscast: A test of grounded language acquisition. In *Proceedings of 25th International Conference on Machine Learning (ICML-2008)*.

Chiang, D. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL-05)*, 263–270.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Perez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence* 89(1-2):31–71.

Feldman, J. A.; Lakoff, G.; Bailey, D. R.; Narayanan, S.; Regier, T.; and Stolcke, A. 1996. L0 – the first five years of an automated language acquisition project. *AI Review* 8.

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 *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, 104–111.

Fleischman, M., and Roy, D. 2007. Situated models of meaning for sports video retrieval. In *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-07)*.

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

Gorniak, P., and Roy, D. 2005. Speaking with your sidekick: Understanding situated speech in computer role playing games. In *Proceedings of the 4th Conference on Artificial Intelligence and Interactive Digital Entertainment*.

Harnad, S. 1990. The symbol grounding problem. *Physica D* 42:335–346.

Kate, R. J., and Mooney, R. J. 2006. Using string-kernels for learning semantic parsers. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL-06)*, 913–920.

Kate, R. J., and Mooney, R. J. 2007. Learning language semantics from ambiguous supervision. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI-2007)*, 895–900.

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

Lodhi, H.; Saunders, C.; Shawe-Taylor, J.; Cristianini, N.; and Watkins, C. 2002. Text classification using string kernels. *Journal of Machine Learning Research* 2:419–444.

Manning, C. D., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Ray, S., and Craven, M. 2005. Supervised versus multiple instance learning: An empirical comparison. In *Proceedings of 22nd International Conference on Machine Learning (ICML-2005)*, 697–704.

Roy, D. K., and Pentland, A. P. 2002. Learning words from sights and sounds: a computational model. *Cognitive Science* 26:113–146.

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

Roy, D. 2005. Grounding words in perception and action: Insights from computational models. *Trends in Cognitive Science* 9(8):389–396.

Sebe, N.; Cohen, I.; Garg, A.; and Huang, T. 2005. *Machine Learning in Computer Vision*. Berlin: Springer Verlag.

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 Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL-06)*, 439–446.

Wong, Y., and Mooney, R. J. 2007a. Generation by inverting a semantic parser that uses statistical machine translation. In *Proceedings of Human Language Technologies: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT-07)*, 172–179.

Wong, Y., and Mooney, R. J. 2007b. Learning synchronous grammars for semantic parsing with lambda calculus. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL-07)*, 960–967.

Wu, D. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics* 23(3):377–403.

Yu, C., and Ballard, D. H. 2004. On the integration of grounding language and learning objects. In *Proceedings of the Nineteenth National Conference 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 *Proceedings of the Thirteenth National Conference on Artificial Intelligence (AAAI-96)*, 1050–1055.

Zettlemoyer, L. S., and Collins, M. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of 21th Conference on Uncertainty in Artificial Intelligence (UAI-2005)*.