Language as part of sensorimotor behavior

Ealan A. Henis Stephen E. Levinson

AT&T Bell Laboratories
600 Mountain Ave. Murray Hill, NJ 07974

Abstract

We present two new systems that learn to understand natural language instructions and carry out implied sensorimotor tasks. These systems constitute a further step in our ongoing investigation of acquisition of language by machines [Henis et al. 1994; Henis et al. 1995]. The research approach is based on the idea that language skills and the understanding of the meaning conveyed by spoken communication should be acquired simultaneously and in conjunction with sensorimotor integration. One system is a simulated manipulator in a three dimensional blocks world. The second is a small mobile robot in an office environment. The systems learn on the basis of semantic-level reinforcement feedback signals provided by the user. The associations of words and sensory inputs are derived on the basis of the mutual information between constituents of the inputs and motor outputs. In the course of interacting with the world, an internal representation of the world is constructed that can be manipulated linguistically. The systems starts off without any task relevant words. It takes about 400 sentences (approximately 40 for the simpler mobile robot) for the system to acquire enough knowledge to respond reasonably to the user input, and relate to its environment in terms of given object names, action names and conjunctions of features. The success of these systems show that the principles which formed the basis for our previous systems scale to more complex systems.

1 Introduction

We observe that language skills and the understanding of the meaning conveyed by spoken communication are acquired simultaneously and in conjunction with sensorimotor integration skills during child development. Noting that the sensorimotor interaction with the world is essential in order to acquire a language, we are motivated to hypothesize that language understanding and sensorimotor behavior are intertwined and should be learned and processed together. On the basis of this guiding principle, we have built a four-action system, that learned to perform sensory based manipulatory actions in response to unrestricted natural language input text in a three-dimensional simulated blocks world [Henis et al. 1994; Henis et al. 1995]. In the current paper we report on two new systems, based on this very same principle. One system extends the action space of the previous simulated system from four to fifteen maneuvering and manipulatory actions. The second system is novel in two respects: it is a *real-world* small mobile robotic system that can perform eight actions (four maneuvering motions and four distinct output sounds), and it receives *speech* input from the user, rather than text input. The input speech was recognized and converted to text by using the BLASR [Zhou et al. 1995] speech recognizer as a front end.

1.1 Background

One possible approach to building an interactive system is to handcraft all the knowledge that we have on the system and its foreseen interactions with the environment. Such knowledge includes physical models of the system/environment and a recipe for the evaluation of alternative actions to be taken, given the occurrence of an event. One drawback of this approach is the enormous amount of information involved. Another drawback is the low fidelity of the predefined models and their limited scope (e.g., [Maes and Brooks 1990]). It is controversial whether such approach could result in goal directed behavior, which requires the activation of combinations of many related actions, perhaps beyond the capabilities of a system based on local rules. Rather than handcrafting our world knowledge into the system, we prefer an alternative approach whereby the system incrementally generates its own internal model of the world in the course of interacting with it.

In order for learning to occur, the system must be provided with feedback. The highest quality feedback is a specification of the correct sequence of actions to be taken, given the current combination of inputs.

Such feedback may be available from a *teacher* (either a human or a device) who possesses knowledge what the appropriate response of the system should be ("supervised learning" [Rummelhart and McClelland 1986]). Since such high quality feedback is not always available, we chose to use reinforcement learning, whereby the system is given only an evaluation of the appropriateness of its overall response (correct/incorrect feedback signals) [Kohonen 1983; Barto and Anandan 1985; Gorin et al. 1991; Sankar and Gorin 1993; Gorin et al. 1994].

The system must learn to "understand" what to do. It is extremely difficult to precisely define "understand". In our work, we used a limited. operational definition of understanding [Gorin et al. 1991] as follows. The machine is said to have "understood" the input sentences in the sensorimotor context if it mapped the combined inputs (sentences and sensory data) into the appropriate action(s) in various environmental situations, as judged by a human observer. Therefore, the machine understands if it performs what *we* expect of it. Clearly this operational definition depends on the world in which the system functions, on the task, and on the structure of the system.

1.2 The Systems

A functional block diagram of the systems is presented in Figure 1. The systems have a structured connectionist architecture embedded in a semantic-level feedback control loop. The information relevant to the input/output periphery of the system (input words, sensed object features, motor outputs) are grouped into categories (e.g., category of words, category of colors, category of roughness values etc.). The associations between members of these categories are learned on the basis of the reinforcement feedback signals. This structured architecture reflects our understanding of the information processing task in general, and of the expected solutions in particular. By incorporating this structure, the system can learn the correct responses in the context of its sensorimotor state more rapidly than when a homogeneous (structure-less) architecture is used [Kohonen 1983; Barto and Anandan 1985; Rummelhart and McClelland 1986; Elman 1991; Plumbley 1991]. Furthermore, this structure provides automatic generalization between objects that share features.

The algorithm that forms the basis for the evolution of associations between entries of the various categories is based on the "mutual information" [Goodman et al. 1988; Plumbley 1991; Sankar and Gorin 1993] between constituents of the inputs (user words and sensed object features) and constituents of the outputs (actions,

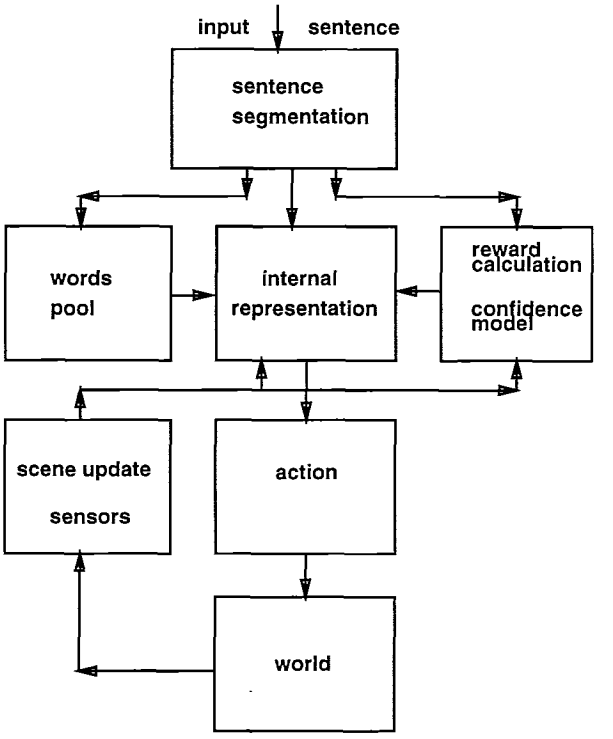

Figure 1: *Functional block diagram.* The systems have a structured connectionist architecture embedded in a semantic-level feedback control loop.

selected object features) as inferred from positively reinforced interactions.

The connection weights which represent the associations captured by the system are derived as follows. Two types of associations are recorded: associations between words and actions/features (e.g., the word "grab" is associated with closing the gripper jaws) and associations between features and other actions/features (e.g., the feature blue may be associated with the shape cylinder and with the action of lifting). The former associations are termed "primary" and the latter associations are termed "secondary". For each object (and action), the primary (and separately the secondary) associations are represented by U_i:

$$U_i = \sum_{feature-j} U_i^j \tag{1}$$

where U_i^j is the contribution of feature category j. For the primary weights, the score of the k-th entry of the $j-$th feature category is given by:

$$a_k^j = sigmoid(\sum_{m=1}^{M} O_m w_{mk}^j + w_k^j) \tag{2}$$

where O_m is the probability of the presence of word m in the current sentence, w_{mk}^j is the association weight between word m and entry k in the $j-$th feature category, and w_k^j is the bias between the word category and entry k of feature category j. For the secondary weights, the score of the k-th entry of the $j-$th feature category is given by:

$$b_k^j = sigmoid(\sum_{feature-i \neq j} \sum_{m \in feature-i} O_{ms}^i w_{mk}^{ij} + w_k^{ij}) \tag{3}$$

where $O_{ms}^i = 1$ if the feature entry m of feature category i belongs to the currently selected object s and 0 otherwise, w_{mk}^{ij} is the secondary weight between entry m of feature i and entry k of feature j, and w_k^{ij} is the bias between the category i and entry k of feature category j. The calculation of the weights and biases is based on numerical estimation of the mutual information:

$$w_{mk}^j = I(c_k^j, v_m) = \log \frac{P(c_k^j|v_m)}{P(c_k^j)} \tag{4}$$

where $I(c_k^j, v_m)$ denotes the mutual information between entry k of feature category j and the word v_m. The primary bias for feature category j is given by:

$$w_k^j = \log P(c_k^j). \tag{5}$$

Each of the secondary weights and biases is computed as follows:

$$w_{sk}^{ij} = I(c_k^j, c_s^i) = \log \frac{P(c_k^j|c_s^i)}{P(c_k^j)} \tag{6}$$

where $I(c_k^j, c_s^i)$ denotes the mutual information between feature entry k of feature category j and feature entry s of feature category i. The secondary bias term for the ij secondary subnetwork is given by:

$$w_k^{ij} = \log P(c_k^j). \tag{7}$$

The conditional probabilities are estimated by counts of successful events:

$$P_1(c_k^j|v_m) = \frac{N(v_m, c_k^j)}{N(v_m)} \in [0, 1] \tag{8}$$

and

$$P_1(c_k^j) = \frac{N(c_k^j)}{N_T} \in [0, 1] \tag{9}$$

where $N(v_m, c_k^j)$ is the number of co-occurrences of entry k of feature category j with word v_m in positively reinforced interactions, $N(v_m)$ is the number of occurrences of word v_m, and N_T is the total number of weight updates. Similarly, for the secondary nets:

$$P_1(c_k^j|c_s^i) = \frac{N(c_s^i, c_k^j)}{N(c_s^i)} \in [0, 1] \tag{10}$$

and

$$P_1(c_k^j) = \frac{N(c_k^j)}{N_T} \in [0, 1]. \tag{11}$$

To improve the approximation for low counts we perform smoothing with a prior belief [Gorin et al. 1991]. The prior belief for $P(c_k^j)$ is

$$P(c_k^j) = \frac{1}{K}. \tag{12}$$

where K is the total number of entries in the feature j. Hence the estimate $P_1(c_k^j)$ is replaced by:

$$P_2(c_k^j) = (1 - \alpha)\frac{1}{K} + \alpha P_1(c_k^j) \tag{13}$$

where $\alpha = \frac{N_T}{m_\alpha + N_T}$ and m_α is the mass of the prior belief. The prior belief for the primary conditional probabilities is:

$$P(c_k^j|v_m) = P(c_k^j) \tag{14}$$

(independent of v_m). Hence the estimate for the conditional probabilities is replaced by:

$$P_2(c_k^j|v_m) = (1 - \beta)P_2(c_k^j) + \beta P_1(c_k^j|v_m) \tag{15}$$

where $\beta = \frac{N(v_m)}{m_\beta + N(v_m)}$. Additional improvement is obtained by clipping the estimates to the prior beliefs unless some threshold is exceeded [Sankar and Gorin 1993].

If there are two or more sentences in a conversation segment, the calculation of a_k^j is repeated for each sentence, and the contributions of all the sentences are averaged with equal weights: After L sentences the score $A_k^j(L)$ is given by:

$$A_k^j(L) = \gamma(L)a_k^j + (1 - \gamma(L))A_k^j(L-1) \qquad (16)$$

where $A_k^j(0) = 0$, a_k^j is the score of the last sentence, and $\gamma(L) = \frac{1}{L}$. The contribution of feature j to the primary networks is:

$$U_i^j = \sum_{entry-k} M(k,i)A_k^j \qquad (17)$$

Where $M(k,i) = 1$ if object i has value k of feature category j. Similarly, for the contribution of feature j to the secondary networks we have:

$$U_i^j = \sum_{entry-k} M(k,i)b_k^j. \qquad (18)$$

In contrast with other connectionist reinforcement learning methods [Kohonen 1983; Barto and Anandan 1985; Rummelhart and McClelland 1986; Elman 1991], the mutual information association weights are derived directly from temporal and spatial correlations that occur throughout the system's history, and uniquely characterize the machine's experience. Furthermore, each weight has an independent meaning (as the mutual information between internal structures that represent external entities), in addition to the collective meaning that emerges from using connectionist networks.

Using the primary associations results in a "content word" approach [Miller and Gorin 1993; Sankar and Gorin 1993] to language acquisition, which is superior to keyword spotting. In this "content word" approach each word is assigned a vector of weights which measure the association between the word and its connotations in terms of the machine's periphery (input/output). Hence, the vector of weights represents a spectrum of responses (rather than one response) for the machine, augmenting the predetermined fixed selections used in keyword-based systems.

At any given time, the system selects one motor action, out of a repertoire of action primitives, to be performed on one selected object. In the simulated system the action list was composed of: move forward, move backward, move left, move right, move up, move down, move left of an object, move right of an object, move above an object, move beneath an object, move in front of an object, move behind an object, attend (move) to an object open the gripper and close the gripper. In the mobile robot system the action list was composed of: move forward, move backward, turn left, turn right and make four different sounds.

The selection of the next action and object are based on time-dependent potential functions associated with each object and action. Each object i has an associated potential function $P_i(t)$, and each action n has an associated potential function $Z_n(t)$. The object $I(t)$ selected at time t is the one that has minimum potential, and similarly for the selected action $N(t)$:

$$I(t) = arg\min_i(P_i(t)); \; N(t) = arg\min_n(Z_n(t)), \qquad (19)$$

where

$$P_i(t) = -B_i(t)U_i(t) + E_i, \qquad (20)$$

$$Z_n(t) = -B_n(t)U_n(t) + E_n + NextError_n + Cost_n. \qquad (21)$$

Here t denotes time, E_i and E_n denote semantic-level error values, and $NextError_n$ is the penalty for following the previous action. The functions $B_i(t)$ and $B_n(t)$ measure the boredom associated with each object and action. These functions decay over time.

$$B_i(t) = \exp(-\lambda_2 \sum_{\tau=t-T_0}^{t} e^{-\lambda_1 d_i(\tau)}), \qquad (22)$$

where d_i is the distance between the object i and the current locus of attention. The parameter T_0 is an averaging time interval. The time-dependent potential functions balance and trade-off the machine's built-in tendencies, and drive its behavior.

The system alternates between two modes. In the user-machine mode, the user is providing natural language sentences that encode a meaning and a goal for the system. The user also provides the machine with feedback: an evaluation of the appropriateness of the most recent response. Three predefined words signal the user's reinforcement: "good", "wrong-action" and/or "wrong-object". All task related words are not predefined, and are learned in the course of the user-machine interactions. In the autonomous mode, the user is not interacting with the machine, and the machine responds to the sensory stimuli on the basis of its current world model (as reflected in the secondary associations) and innate properties. The latter are a balance between lock of attention on an object, and the time-dependent boredom function associated with each object. These innate tendencies drive the machine both to inspect an object and to explore new parts of its environment by trying to execute new actions.

1.3 Results

The system started off with no a priori known vocabulary words, and with simulated visual and tactile in-

formation about the objects in the scene (e.g., their color, shape, height, roughness etc.). The machine's responses to the first few sentences were typically inappropriate. These initial responses were based on still too weak associations between words/sensory-inputs and output-features/actions. Following the negative feedback signals supplied by the user, the machine selected other responses, and eventually performed an appropriate action on the correct object. After the appropriate response was obtained, and following the positive reinforcement signals, the relevant associations were strengthened. As the interaction continued, a larger percentage of the machine's responses were acceptable and were judged appropriate. It took about 4 hours of interaction (400 sentences) for the 15-action simulated system to arrive at a state in which the machine correctly responded to a high percentage of the user's instructions within 1-2 input sentences. Since the action-space of the mobile robotic system is much smaller than that of the simulated system, it took only 15 minutes to arrive at a similar state. Following this initial phase, the machine learned the correct associations between words and the relevant internal structures, and subsequently appropriately responded to most of the user's input, giving users the subjective feeling that the machine "understands" their instructions. Several users made the machine rearrange the objects in the scene as they pleased by issuing sequences of natural language instructions. Thus, the machine has demonstrated its potential to learn to appropriately relate to its world in terms of conjunctions of object features and given object names, and to perform the motor actions that are implied by the user in the sensory context.

2 Discussion

Language conveys messages that communicate meaning. The primary goal of the systems presented above was to decode the intentions of the user and respond accordingly, in the context of the sensorimotor state of the machines. Extracting the meaning of natural language is a formidable task, even within a limited semantic domain. The machine's performance is not close to that of a human. Nevertheless, rapidly the system learned to performs many useful instructives, without a predefined linguistic model. The structured architecture allows to generalize between objects that share a feature. The principle that language understanding is part of behavior and must be learned and processed in conjunction with sensorimotor behavior has proven to be useful.

How would the system scale to more complex action spaces? As we add actions and objects having various features to the scene, the number of possible output selections becomes larger. Although the performance of the system following training will not substantially degrade, the amount of trial and error interactions will rapidly increase and will stretch the user's patience. If the system is extended further to allow using parameterized actions, the current scanning method (specifying "wrong-object" and "wrong-action" when the machine errs) would fail altogether, and must be replaced with some other (perhaps grey-level) feedback signals.

To improve the language understanding capabilities of the machine in the future, the currently used semantic primitives (words) may be replaced by ordered word-pairs and triplets, allowing for better context dependent language understanding. Another possibility is to use a phonetic transcription of the input sounds, and to let the machine learn the notion of semantic units from the input phonetic stream. This approach will put to test the idea that the semantic units of language (words) emerge from the continuous stream of heard sounds via the detection of correlations between the latter and sensorimotor inputs/outputs. Other extensions of the current system in the future might include using real-time sensory inputs, instead of the simulated ones. Constructing such systems will enable us to gain insight into processes that might allow machines to learn to communicate and interact with the environment in a human-like fashion.

References

[1] A. Barto and P. Anandan, "Pattern recognizing stochastic learning automata". *IEEE Trans. on Systems, Man and Cybernetics, vol. SMC-15, no. 3*, 360 - 375, 1985.

[2] J. L. Elman "Distributed representations, simple recurrent networks, and grammatical structure". *Machine Learning vol. 7 no. 2/3*, 195 - 226, 1991, (special issue on connectionist approaches to language learning).

[3] A. N. Gertner and A. L. Gorin, "Adaptive Language Acquisition for an Airline Information Subsystem", *Artificial Neural Networks for Speech and Vision* 401-428, (ed. R. Mammone), Chapman and Hall, 1993.

[4] R. M. Goodman J. W. Miller and P. Smyth, "An information theoretic approach to rule-based connectionist expert system". *Advances in neural information processing systems I*, D.S. Touretzky

(ed.), 256 -263, Morgan Kaufmann, San Mateo Ca, 1988.

[5] A. L. Gorin S. E. Levinson, A. N. Gertner and E. Goldman, "Adaptive acquisition of language". *Computer Speech and Language Vol. 5, no. 2*, 101 - 132, 1991.

[6] A. L. Gorin, S. E. Levinson and A. Sankar, "An experiment in spoken language acquisition". *IEEE Trans. on Speech and Audio. Vol. 2, 40.1, part II*, 224 - 240, 1994.

[7] E. A. Henis, S. E. Levinson and A.L. Gorin. "Mapping natural language and sensory information into manipulatory actions." *Proceedings of the Eighth Yale Workshop on Adaptive and Learning Systems* 190-195. Yale University, New Haven CT, June 1994.

[8] E. A. Henis, S. E. Levinson and A.L. Gorin. "Mapping natural language and sensory information into manipulatory actions". Submitted to *International Journal of Pattern recognition and Artificial Intelligence*, 1995.

[9] T. Kohonen, *Self organization and associative memory.* Springer series in information sciences 8, Springer-Verlag Berlin Heidelberg, 1983.

[10] P. Maes and R. A. Brooks, "Learning to coordinate behaviors". *AAAI-90*, 796 - 802, 1990.

[11] B. Mel, *Connectionist robot motion planning: a neurally inspired approach to visually guided reaching.* Academic Press, San Diego, CA, 1991.

[12] L. G. Miller and A. L. Gorin, "Structured networks for adaptive language acquisition", *Int. J. of Pattern Recognition and Artificial Intelligence Vol 7 No. 4*, 873 - 898, 1993.

[13] M. D. Plumbley, "On information theory and unsupervised neural networks." *Ph.D. Dissertation*, Cambridge University, UK, 1991.

[14] D. E. Rummelhart and J. L. McClelland (eds.), *Parallel distributed processing*, MIT Press, 1986.

[15] M. Salganicoff and R. Bajcsy, "Sensorimotor learning using active perception in continuous domains". *University of PA Dept. of CS technical report MS-CIS-91-87 GRASP Lab 284*, 1991.

[16] A. Sankar and A. L. Gorin, "Visual focus of attention in adaptive language acquisition". In *Artificial Networks for Speech and Vision Applications*, 357 - 387, R. Mammone (ed.), Chapman and Hall, 1993.

[17] A. Sankar, A. L. Gorin, J. G. Wilpon, S. Y. Lee, R. Venkataramani and D. E. Bock, "Language acquisition for automated call routing in a telephone network". *Bell Labs technical memorandum*, 1993.

[18] C. Shannon, "A mathematical theory of communication". *Bell System Technical Journal vol. 27 no. 3*, 379 - 423, 1948.

[19] T. Winograd, "What does it mean to understand language". *Cognitive Sci. vol. 4*, 209-241, 1980.

[20] Q. Zhou, W. Chou, F. Chen and R. Ritenour, "BLASR-RT Toolkit for interactive voice response applications." *AT&T Bell Laboratories Technical Memorandum*, 1995.