A Computational Model of Embodied Language Acquisition

Content area keywords:
cognitive modeling, machine learning, perception

Abstract

We present a computational model which simulates the early stages of human language acquisition by associating words with their perceptually grounded meanings. The central idea is to explore the role of attention in lexical learning and make use of body movements (as deictic references) to discover temporal correlations of data from different modalities. The implemented model can be trained in an unsupervised manner in which users perform everyday tasks while providing natural language descriptions of their behaviors. The system is embodied by sharing user-centric multisensory information consisting of acoustic signals, user's perspective video, gaze positions, head directions and hand movements. A multimodal learning algorithm uses these data to firstly spot words from continuous speech and then associate action verbs and object names with their perceptually grounded meanings.

1 Introduction

Humans develop based on their sensorimotor experiences with the physical environment. Different levels of abstraction are necessary to efficiently encode those experiences, and one vital role of human brain is to bridge the gap from embodied experience to its expression as abstract symbols. To mimic human perceptual abilities, a challenge in machine intelligence is how to establish a correspondence between symbolic representations in an intelligent system situated in the physical world (e.g., a robot or an embodied agent) and sensor data collected from the environment, which is termed as the symbol grounding problem by [Harnad, 1990].

Language is about symbols and humans ground those symbols in sensorimotor experiences during their development [Lakoff and Johnson, 1980]. This paper explores the grounding problem by studying and computationally modeling how humans ground semantics, which is the key to understanding our own minds and ultimately create artificial ones. To do so, it is helpful to make use of our knowledge of human language development to guide the design of our approach. For infants, around 6 to 12 months is the stage of grasping the first words. A predominant proportion of most children's first vocabulary (the first 100 words or so), in various languages and under varying child-rearing conditions, consist of object names, such as food, clothing and toys. The second large category is the set of verbs that is mainly limited to action terms. Gillette et al. [Gillette *et al.*, 1999] showed that learnability of a word is primarily based upon its imageability or concreteness. Therefore, most object names and action verbs are learned before other words because they are more observable and concrete. Next, infants move to the stage of vocabulary spurt or naming explosion, in which they start learning large amounts of words more rapidly than before. At the meanwhile, grammar gradually emerges from the lexicon, both of which share the same mental-neural mechanisms [Bates and Goodman, 1999]. Many of the later learned words correspond to abstract notions (e.g., noun:"idea", verb:"think") and are not directly grounded in embodied experience. However, Lakoff and Johnson [Lakoff and Johnson, 1980] explained that all human understanding is based on metaphorical extension of how we perceive our own bodies and their interactions with the physical world. Thus, the initial and imageable words directly grounded in physical embodiment serve as a foundation for the acquisition of abstract words and syntax which become indirectly grounded through their relations to those grounded words. Therefore, the initial stage of language acquisition, in which infants deal primarily with the grounding problem, is critical in this semantic bootstrapping procedure because it provides a sensorimotor basis for further development.

Inspired by infant language acquisition, we develop an unsupervised learning mechanism that can learn perceptually grounded meanings of words from users' everyday activities. The only requirement is that users need to provide natural language descriptions of their behaviors while performing those day-to-day tasks. To learn a word (shown in Figure 1), the system needs to discover its sound pattern from continuous speech, identify its possible meaning from extralinguistic context, and associate these two. The range of problems we need to address in this kind of unsupervised word learning is substantial, so to make concrete progress, this paper focuses on how to associate visual representations of objects with their spoken names and map body movements to action verbs. Our work suggests a new kind of embodied learning system in machine intelligence which can automatically acquire perceptually grounded lexicons by being situated in our everyday lives and sharing user-centric multisensory data.

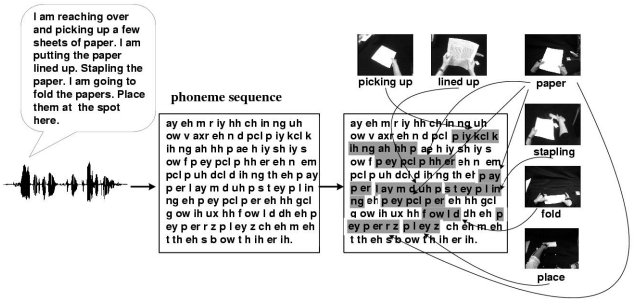

Figure 1: **The problems in word learning.** The raw speech is firstly converted to phoneme sequences. The goal of our method is to discover phoneme substrings that correspond to the sound patterns of words and then infer the meanings of those words from non-speech modalities.

2 The Role of Attention in the Embodiment of Language

The topics related to language acquisition, such as speech segmentation and word learning, have attracted many attentions in both psycholinguistics (e.g., [Saffran *et al.*, 1996]; [Tomasello, 2001]) and computer science (e.g., [Siskind, 1996]; [Steels and Vogt, 1997]; [Roy and Pentland, 2002]). However, most computational studies of language acquisition (for a review, see [Brent, 1999]) focus on the role of linguistic information as the central constraint. Extralinguistic information, such as visual perception and speaker's attention, is ignored although it also plays a critical role in human language acquisition. Recently, a new trend termed *associationism* assumes that language acquisition is solely based on statistical learning of co-occurring data from linguistic modality and extralinguistic context. In [Saffran *et al.*, 1996], Saffron et al. have provided experimental evidences that both children and adults are sensitive to features of co-occurrence statistics in language. In [Roy and Pentland, 2002], they have proposed and implemented a computational model of infant language learning, in which they used the temporal correlation of speech and vision to associate spoken utterances with a corresponding object's visual appearance. It seems quite reasonable to assume that human cognitive system exploits those statistical information. However, despite the merit of this idea, *associationism* is unlikely to be the whole story because it is based on the assumption that words are always uttered when their referents are perceived, which has not been verified by experimental studies of infants [Bloom, 2000].

In addition to temporal co-occurrences of multisensory data, recent psycholinguistic studies (e.g., [Bloom, 2000]; [Tomasello, 2001]) have shown that the other major source of constraints in language acquisition is social cues, such as eye gaze and gesture pointing, which can be utilized to infer speakers' referential intentions. Bloom [Bloom, 2000] argued that children's word learning actually draws extensively on their understanding of the thoughts of speakers, which is considered as a precursor to lexical development. His claim has been supported by the experiments in which young children were able to figure out what adults were intending to refer to by speech. In a complementary study of embodied cognition [Ballard *et al.*, 1997], Ballard and colleagues proposed a cognitive system of implicit reference termed deictic, in which the body's pointing movements are used to bind objects in the world to variables in cognitive programs of our brain.

These cognitive studies suggest the value of a formal model that explores the computational role of attention in lexical development. In the model described in the next section, a speaker's focus of attention is estimated and utilized to facilitate lexical learning in two ways. Firstly, *attentional contexts* composed of attentional objects and intentional actions provide cues for word spotting from a continuous speech stream. Speech segmentation without prior language knowledge is a challenging problem and has been addressed by solely using linguistic information. In contrast, our method appreciates the importance of extralinguistic context in which spoken words are uttered. We propose that the sound patterns frequently appearing in the same attentional context are likely to have grounded meanings related to this context. Thus, by finding frequently uttered sound patterns in a specific context (e.g., an attentional object that users look toward or an intentional action that users perform), the model discovers word-like sound units as candidates for building lexicons. Secondly, a difficult task of word learning is to figure out which entities specific words refer to from a multitude of co-occurrences between words and things in the world. This is accomplished in our model by utilizing user's intentional body movements as deictic references to establish associations between spoken words and their perceptually grounded meanings.

3 An Embodied Language Acquisition System

In order to ground semantics, a computational system needs to have sensorimotor experiences by interacting with the physical world. To do this we attach different kinds of sensors to a real person as shown in Figure 2. Those sensors include a head-mounted CCD camera to capture a first-person point of view, a microphone to sense acoustic signals, an eye tracker to track the course of eye movements that indicate the agent's attention, and position sensors attached to the head and hands of the agent to simulate proprioception in sense of motion. The functions of those sensors are similar to human sensory systems and they allow our computational system to collect user-centric multisensory data so that it sees as the agent sees, hears as the agent hears and experiences the life in the first-person sense.

Figure 2: The intelligent system shares user-centric multisensory information with a real agent. This allows the association of coincident signals in both linguistic modality and extralinguistic(contextual) modalities.

In typical scenarios, an agent performs everyday tasks while describing his/her actions verbally. To learn words from the user's spoken descriptions, three fundamental problems needed to be addressed are: (1) action recognition and object recognition to identify grounded meanings of words encoded in non-speech contextual information, (2) speech segmentation and word spotting to extract the sound patterns of the individual words which might have grounded meanings, (3) association between spoken words and their meanings.

3.1 Clustering Perceptually Grounded Meanings

The non-linguistic inputs of the system consist of visual data from a head-mounted camera, head and hand positions in concert with gaze-in-head data. Those data provide attentional contexts in which spoken utterances are produced. Thus, the possible meanings of spoken words that users utter are encoded in those contexts, and we need to extract those meanings from raw sensory inputs. Specifically, the system should spot and recognize actions from user's body movements, and discover the objects of user interest. As a result, we will obtain two temporal sequences of grounded meanings as depicted by the box labeled "attentional contexts" in Figure 3.

In accomplishing well-learned tasks, a user's focus of attention is linked with body movements. In light of this, our method firstly utilizes eye and head movements as cues to estimate the user's focus of attention. Attention, as represented by gaze fixation, is then used for spotting the target object of user interest. Specifically, at every attentional point in time, we make use of eye gaze as a seed to find the attentional object from all the objects in a scene. Then the object is represented by a feature vector consisting of color, shape and texture features. For further information see the method developed by [Yu *et al.*, 2002]. Next, since the feature vectors extracted from visual appearances of attentional objects do not occupy a discrete space, we vector quantize them into clusters by applying a hierarchical agglomerative clustering algorithm. Finally, for each cluster we select a prototype to represent perceptual features of this cluster.

Meanwhile, attention switches are calculated based on eye and head movements, which are then used to segment a sequence of hand movements into action primitives. For each segment, a time series of feature vectors are extracted from hand positions. Further information about action segmentation can be obtained from the work of [Yu and Ballard, 2002b]. Now we want to cluster these time series into groups so that each group corresponds to a qualitatively different type of motion. This goal is achieved by applying a Dynamic Time Warping (DTW) algorithm to calculate the distances of time series and then clustering them based on those distances. In this way, we can quantize time series into discrete symbols corresponding to clusters. Cluster centroids are computed and selected as prototypes to represent different actions.

3.2 Speech Processing

We describe our methods of phoneme recognition and phoneme string comparison in this subsection, which provide a basis for further processing. Detailed descriptions of algorithms can be obtained from the work of [Yu and Ballard, 2002a].

Phoneme Recognition

An endpoint detection algorithm segments a speech stream into several spoken utterances. Then the speaker independent phoneme recognition system developed by Robinson [Robinson, 1994] is employed to convert spoken utterances into phoneme sequences. The method is based on Recurrent Neural Networks (RNNs) that perform the mapping from a sequence of the acoustic features extracted from raw speech to a sequence of phonemes. We trained RNNs using TIMIT database — phonetically transcribed American English speech — which consists of read sentences spoken by 630 speakers from eight dialect regions of the United States. Once trained, a dynamic programming match is made to find the most probable phoneme sequence of a spoken utterance, for example, the boxes labeled "phoneme strings" in Figure 3.

Comparing Phoneme Sequences

Our comparison of phoneme sequences has two purposes: one is to find the longest similar substrings of two phonetic sequences (word-like units spotting described in Subsection 3.3), and the other is to cluster segmented utterances represented by phoneme sequences into groups (word-like units clustering presented in Subsection 3.3). Given raw speech input, the specific requirement here is to cope with the acoustic variability of spoken words in different contexts and by various talkers. Due to this variation, the outputs of the phoneme recognizer are noisy phoneme strings that are different from phonetic transcriptions of text. Thus, the goal of phonetic string matching is to identify sequences that might be different actual strings, but have similar pronunciations.

To align phoneme sequences, we first need a metric for measuring distances between phonemes. A phoneme is represented by a 15-dimensional binary vector in which every entry stands for a single articulatory feature called a distinctive feature. Those distinctive features are indispensable attributes of a phoneme that are required to differentiate one phoneme from another in English. We compute the distance between two individual phonemes as the Hamming distance. Based on this metric, a modified dynamic programming algorithm is developed to compare two phoneme strings by measuring their similarity. A similarity scoring scheme assigns large positive scores to pairs of matching segments, large negative scores to pairs of dissimilar segments, and small negative scores to the operations of insertion and deletion to convert one sequence to another. The optimal alignment is the one that maximizes the accumulated score. See the work of [Yu and Ballard, 2002a] for further information about the method of phoneme sequence comparison.

3.3 Word Learning

This subsection describes our approach to integrating multimodal data for word acquisition. We divide this problem into two basic steps: speech segmentation shown in Figure 3 and lexical acquisition illustrated in Figure 5.

Word-like Unit Spotting

Figure 3 illustrates our approach to spotting word-like units in which the central idea is to utilize attentional contexts to facilitate word spotting. The reason we use the term "word-like units" is that some actions are verbally described by verb

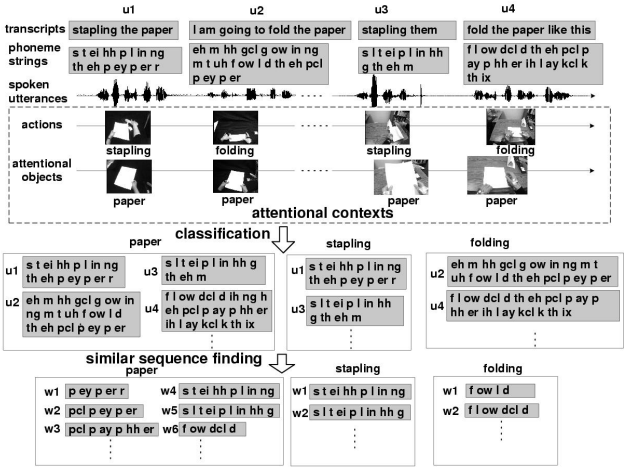

Figure 3: **Word-like unit segmentation**. Spoken utterances are categorized into several bins that correspond to temporally co-occurring actions and attentional objects. Then we compare any pair of spoken utterances in each bin to find the similar subsequences that are treated as word-like units.

phrases (e.g., "line up") but not single action verbs. The inputs are phoneme sequences (i.e. u_1, u_2, u_3, u_4) and possible meanings of words (objects and actions) extracted from the environment. Those phoneme utterances are categorized into several bins based on their possibly associated meanings. For each meaning, we find the corresponding phoneme sequences uttered in temporal proximity, and then categorize them into the same bin labeled by that meaning. For instance, u_1 and u_3 are temporally correlated with the action "stapling", so they are grouped in the same bin labeled by the action "stapling". Since one utterance could be temporally correlated with multiple meanings grounded in different modalities, it might be selected and classified in different bins. For example, the utterance "stapling a few sheets of paper" is produced when a user performs the action of "stapling" and looks toward the object "paper". In this case, the utterance is put into two bins: one corresponding to the object "paper" and the other labeled by the action "stapling". Next, based on the method described in Subsection 3.2, we compute the similar substrings between any two phoneme sequences in each bin to obtain word-like units. Figure 4 shows an example of extracting word-like units from the utterance u_2 and u_4, both of which are in the bin of the action "folding".

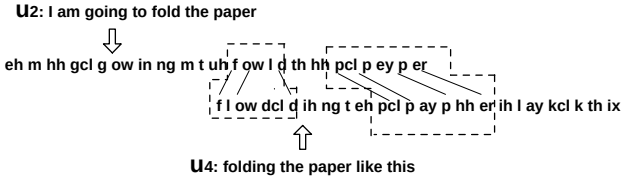

Figure 4: **An example of word-like unit spotting**. The similar substrings of two sequences are /f ow l d/ (fold), /f l ow dcl d/ (fold), /pcl p ey p er/ (paper) and /pcl p ay p hh er/ (paper).

Word-like Unit Clustering

As shown in Figure 5, the extracted phoneme substrings of word-like units are clustered by a hierarchical agglomerative clustering algorithm that is implemented based on the method described in Subsection 3.2. The centroid of each cluster is then found and adopted as a prototype to represent this cluster. Those prototype strings are associated with their possible grounded meanings to build hypothesized lexical items (shown in Figure 5). Among them, some are correct ones, such as /s t ei hh p l in ng/ (stapling) associated the action of "stapling", and some are incorrect, such as /s t ei hh p l in ng/ (stapling) paired with the object "paper". Now that we have hypothesized word-meaning pairs, the next step is to select correct lexical items.

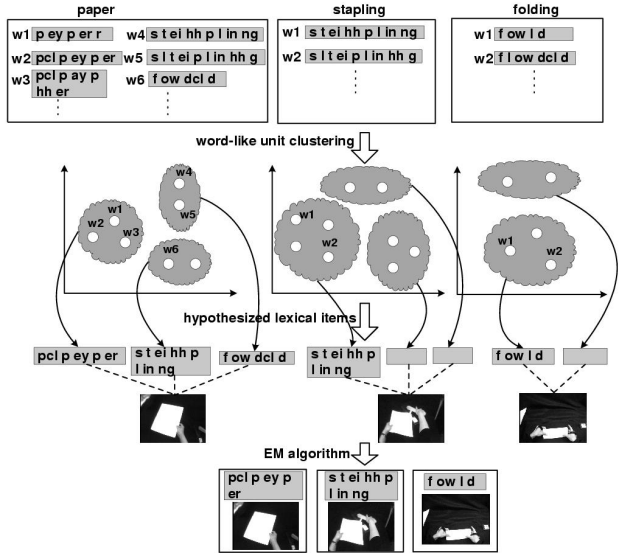

Figure 5: **Word learning.** The word-like units in each bin are clustered based on the similarities of their phoneme strings. The EM-algorithm is applied to find lexical items from hypothesized word-meaning pairs.

Multimodal Integration

We utilize the co-occurrence of multimodal data to build semantic representations that associate visual features of objects and body movements with spoken words. We take a novel view of this problem as the word correspondence problem in machine translation. For example, body movements can be looked as a "body language". Thus, associating body movements with action verbs can be viewed as the problem of identifying word correspondences between English and "body language". In light of this, we apply a technique from machine translation to address this problem. We model the probability of each word as a mixture model that consists of the conditional probabilities of each word given its possible meanings. In this way, the Expectation-Maximization(EM) algorithm is employed to find the reliable associations of spoken words and their grounded meanings that maximize the probabilities.

We assume that every meaning m can be associated with a word-like phoneme string w. We can find the word $\tilde{w}$ that is associated with the meaning m by choosing the one that maximizes $P(w|m)$. Let N be the number of meanings, W_n be the number of words in the n-th meaning, and let a_n represent a set of the possible assignments:$(a_{n1}, a_{n2}, ..., a_{nW_n})$, such that a_{nj} assigns the word w_{nj} to the meaning m_n. $p(a_{nj})$ is the probability that the meaning m_n is associated with a specific word w_{nj} and $p(w_{nj}|m_n)$ is the probability of obtaining

the word w_{nj} given the meaning m_n. We use the model similar to that of Duygulu et al. [Duygulu *et al.*, 2002]:

$$p(w|m) = \prod_{n=1}^{N} \prod_{j=1}^{W_n} p(a_{nj}) p(w_{nj}|m_n) \quad (1)$$

We can estimate $p(w_{nj}|m_n)$ from data directly and the only incomplete data is $p(a_{nj})$. The remaining problem is to find the maximum likelihood parameter:

$$\widetilde{p}(a) = \arg\max_{p(a)} p(w|m, p(a)) = \arg\max_{p(a)} \sum_{a} p(a, w|m, p(a)) \quad (2)$$

The EM algorithm can be expressed in two steps. Let $p(a)^{[k]}$ be our estimate of the parameters at the kth iteration. In **E-step**: we compute the expectation of the log-likelihood function:

$$\begin{aligned} Q(p(a)|p(a)^{[k]}) &= E[log\ p(a, w|m, p(a)^{[k]})] \\ &= \sum_{n=1}^{N} \sum_{j=1}^{W_n} p(a_{nj}|w_{nj}, m_n, p(a)^{[k]}) \times \\ &\quad log[p(a_{nj}) p(w_{nj}|m_n)] \end{aligned} \quad (3)$$

In **M-step**: we wish to find assignment probabilities $p(a)^{[k+1]}$ so as to maximize $Q(p(a)|p(a)^{[k]})$ subject to the constraints that for each m_n:

$$\sum_{j=1}^{W_n} p(a_{nj}) = 1 \quad (4)$$

Therefore, we introduce Lagrange multipliers λ_n and seek an unconstrained maximization:

$$h(p(a), \lambda) = Q(p(a)|p(a)^{[k]}) + \sum_{n=1}^{N} \lambda_n (1 - \sum_{j=1}^{W_n} p(a_{nj})) \quad (5)$$

We compute derivatives with respect to the multipliers λ and the parameters $p(a)$ to estimate $p(a_{nj})$:

$$p(a_{nj}) = \frac{p(a_{nj}|w_{nj}, m_n, p(a)^{[k]}) p(w_{nj}|m_n)}{\sum_{j=1}^{W_n} p(a_{nj}|w_{nj}, m_n, p(a)^{[k]}) p(w_{nj}|m_n)} \quad (6)$$

The algorithm sets an initial $p(a)^0$ to be flat distribution and performs the E-step and the M-step successively until convergence. Then for each meaning m_n, the system selects all the words with the probability $p(a_{nj})$ greater than a pre-defined threshold. Thus, one meaning can be associated with multiple words. This is because people may use different names to refer to the same object and the spoken form of an action verb can be expressed differently. In this way, the system is developed to learn all the spoken words that have high probabilities in association with a meaning.

The output of the learning system is the lexicon L, which is defined as a set of word-meaning associations: $L = \{l_i\}$ where $l_i =< w_i, m_i, p_i >$ is a lexical item, in which w_i is a phoneme sequence of a word, m_i is a sensorimotor representation of a meaning (e.g., a feature vector of a visual object or a time series of feature vectors of motion), and p_i is the association score that indicates the certainty of a lexical item.

4 Experiment

We collected data from multiple sensors with timestamps. A Polhemus 3D tracker was utilized to acquire 6-DOF hand and head positions at $40Hz$. A user wore a head-mounted eye tracker from Applied Science Laboratories(ASL). The headband of the ASL held a miniature "scene-camera" to the left of the user's head, which provided the video of the scene from the first-person perspective. The video signals were sampled at the resolution of 320 columns by 240 rows of pixels at the frequency of 15Hz. The gaze positions on the image plane were reported at the frequency of $60Hz$. The acoustic signals were recorded using a headset microphone at a rate of 16 kHz with 16-bit resolution.

Six users participated in the experiments. They were asked to sit at a table and performed the task of "stapling papers" while describing their actions verbally. Each user performed the task six times. Figure 6 shows snapshots captured from the head-mounted camera when a user performed the task.

(a) (b) (c) (d) (e) (f)

Figure 6: The snapshots of an action sequence when a user performed the task of stapling several sheets of paper: (a) picking up papers (b) placing them to the position close to the body (c) lining up (d) stapling (e) folding (f) placing them to the target location.

To evaluate experimental results, we define the following three measures: (1) **Semantic accuracy** is to measure the performance of processing non-linguistic data, which consists of clustering both human actions and visual attentional objects. (2) **Speech segmentation accuracy** is to measure whether the beginning and the end of phoneme strings of word-like units are correct word boundaries. (3) **Word learning accuracy** is to measure the percentage of successfully segmented words that are correctly associated with their meanings.

Table 1: Results of word acquisition

	semantics	speech segmentation	word learning
overall	90.6%	70.6%	89.5%
picking up	95.5%	72.5%	86.9%
placing	92.3%	66.9%	90.2%
lining up	69.2%	70.3%	83.6%
stapling	85.2%	70.6%	83.2%
folding	89.3%	69.8%	85.6%
paper	96.7%	70.8%	92.1%

Table 1 shows the results of three measures. The recognition rate of the phoneme recognizer we used is 75% because it does not encode any language model and word model. Based on this result, the overall accuracy of speech segmentation is 70.6%. Naturally, an improved phoneme recognizer based on a language model would improve the overall results, but the intent here is to study the model-independent learning approach. The error in word learning is mainly caused by a few words (e.g., "several" and "here") that frequently occur in some contexts but do not have grounded meanings. Considering that the system processes raw sensory data, and our learning method works in unsupervised mode without man-

ually encoding any linguistic information, the accuracies for both speech segmentation and word learning are impressive.

5 Conclusions

We have presented a computational model of embodied language learning. Compared to previous work, the novelty of our approach arises from the following three aspects. First, our model shares user-centric multisensory information with a real agent and grounds semantics directly from egocentric experience. Second, both words and their perceptually grounded meanings are acquired from sensory inputs. Furthermore, grounded meanings are represented by perceptual features but not abstract symbols, which provides a sensorimotor basis for machines and people to communicate and understand each other through language. Third, from the perspective of machine learning, both clustering multisensory data and finding correspondences between words and meanings are accomplished in unsupervised mode without additional teaching signals. This is analogical to human development in natural environments which also do not contain manually labeled signals for sensory data. The important information the environments contain is encoded in the temporal correlations between sensations to different sensory modalities, and human language learners make use of this correlational structure. To simulate human counterparts, our approach has shown that by utilizing attentional contexts and deictic references of users' body movements as a substrate to process multimodal information, machines can also learn lexical items without labeled data. Thus, our work suggests a new kind of embodied learning system that can automatically acquire perceptually grounded lexicons by being situated in our everyday lives and sharing user-centric multisensory information.

In our current implementation, clustering individual unimodal data is firstly accomplished to obtain abstract symbolic representations. We then merge those symbolic representations from different modalities to find word-meaning associations and map those symbols to the corresponding sensorimotor representations to build grounded lexicons. This ignores the point that at the early stage, the classification abilities of infants in visual, auditory and proprioceptive subsystems have not been well developed yet and intersensory data can affect other unimodal perceptions by *cross-modal influence* [Coen, 2001]. Thus, in order to make the model more cognitively plausible for simulating human development, it will be necessary to develop a multimodal learning algorithm in which the sharing of perpetual information from different modalities must also occur at the earliest stages of sensory processing.

References

[Ballard *et al.*, 1997] Dana H. Ballard, Mary M. Hayhoe, Polly K. Pook, and Rajesh P. N. Rao. Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20:1311–1328, 1997.

[Bates and Goodman, 1999] Elizabeth Bates and Judith C. Goodman. On the emergence of grammar from the lexicon. In Brian Mac Whinney, editor, *The Emergence of Language*. Lawrence Erlbaum Associates, 1999.

[Bloom, 2000] Paul Bloom. *How children learn the meanings of words*. The MIT Press, Cambridge, MA, 2000.

[Brent, 1999] Michael R. Brent. Speech segmentation and word discovery: A computational perspective. *Trends in Cognitive science*, 3(8):294–301, 1999.

[Coen, 2001] Michael H. Coen. Multimodal integration - a biological view. In *Proceeding of The Seventeenth International Joint Conference on Artificial Intelligence (IJCAI 2001)*, Seattla, WA, 2001.

[Duygulu *et al.*, 2002] P. Duygulu, K. Barnad, J.F.G. de Freitas, and D.A. Forsyth. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *Proceedings of European Conference on Computer Vision*, Copenhagen, 2002.

[Gillette *et al.*, 1999] J. Gillette, H. Gleitman, L. Gleitman, and A. Lederer. Human simulations of vocabulary learning. *Cognition*, 73:135–176, 1999.

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

[Lakoff and Johnson, 1980] G. Lakoff and M. Johnson. *Metaphors We Live By*. University of Chicago Press, Chicago, 1980.

[Robinson, 1994] Tony Robinson. An application of recurrent nets to phone probability estimation. *IEEE Transactions on Neural Networks*, 5(2):298–305, 1994.

[Roy and Pentland, 2002] Deb Roy and Alex Pentland. Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146, 2002.

[Saffran *et al.*, 1996] Jenny R. Saffran, Elissa L. Newport, and Richard N. Aslin. Word segmentation: The role of distributional cues. *Journal of memory and language*, 35:606–621, 1996.

[Siskind, 1996] Jeffrey Mark Siskind. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61:39–61, 1996.

[Steels and Vogt, 1997] Luc Steels and P. Vogt. Grounding adaptive languge game in robotic agents. In *Proc. of the 4th European Conference on Artificial Life*, 1997.

[Tomasello, 2001] Michael Tomasello. Perceiving intentions and learning words in the second year of life. In Michael Tomasello and Elizabeth Bates, editors, *Language Development: The Essential Readings*. Blackwell Publisher, 2001.

[Yu and Ballard, 2002a] Chen Yu and Dana H. Ballard. A computational model of embodied language learning. Technical Report 791, Department of Computer Science, University of Rochester, 2002.

[Yu and Ballard, 2002b] Chen Yu and Dana H. Ballard. Learning to recognize human action sequences. In *IEEE Proceedings of the 2nd International Conference on Development and Learning*, pages 28–34, Boston, MA, 2002.

[Yu *et al.*, 2002] Chen Yu, Dana H. Ballard, and Shenghuo Zhu. Attentional object spotting by integrating multisensory input. In *IEEE Proceedings of the 4th International Conference on Multimodal Interface*, Pittsburg, PA, 2002.