Information Theory and Representation in Associative Word Learning

Brendan Burns, Charles Sutton, Clayton Morrison, Paul Cohen
University of Massachusetts Amherst, Amherst MA 01002
{bburns,casutton,clayton,cohen}@cs.umass.edu

Abstract

A significant portion of early language learning can be viewed as an associative learning problem. We investigate the use of associative language learning based on the principle that words convey Shannon information about the environment. We discuss the shortcomings in representation used by previous associative word learners and propose a functional representation that not only denotes environmental categories, but serves as the basis for activities and interaction with the environment. We present experimental results with an autonomous agent acquiring language.

1. Introduction

All language learning is concerned with the proper use of words and phrases. Early first-language learning is particularly focused on learning words that can be used to refer to the immediate environment (objects, relations, events and processes) or used in simple social activities (expressions, requests, etc.). The early first-language learner is faced with the challenge of determining both what the usable units of language are (words and eventually grammatical constructs) as well as the appropriate contexts of their use. We propose that a significant portion of this early language learning problem can be viewed as an associative learning problem, where the language learner is learning the associations between represented contexts for use (perceptual categories, representations of event structure, etc.) and verbal patterns (discovering word boundaries and the ordered structure of words in sentences). Furthermore, linguistic behavior observed by the learner – the data available to it – conveys information, in Shannon's information-theoretic sense, about contexts of use (as perceptual states, internal states, and representations of processes) and this information-bearing property may be exploited by an associative learner.

Associative learning is the acquisition of meanings through the observation of the co-occurence of the words and an example of their meaning. We define associative learner as one which learns a map $f : W \rightarrow \Phi$, where W is a set of words and phrases, and Φ is the set of the learner's perceptual primitives.

Associative learning is a natural technique for mapping words to their meanings because it does not require the learning agent to have any *a priori* knowledge of the world save the ability to form associations. Associative learning may not explain all language learning, but it is a necessary tool to bootstrap later, more sophisticated lexical development.

The representation of both words and meaning greatly influences what can be learned. Previous associative word learning systems (Section 2.) have mapped individual words to their meanings. However, very few words have meanings which are not dependent upon their larger context in a phrase or sentence. To broaden the scope of the meanings we might learn, and in contrast to other learners which mapped single word-tokens to meanings, our learner considers the meanings of words in context (Section 3.2). Previous associative learners have also mapped words to their denotational meanings, that is, a description of raw sensor input. Although denotational meanings can be used for passive recognition, the learner cannot use the meanings for performing actions. We are interested in constructing learning systems capable of autonomous behavior and active learning. Such a learner must learn *functional* meanings (Section 3.) allowing it to model a word's meaning. Functional meanings are motivated by the notion of a word's functional semantics, and work in cognitive-psychology which indicates that mental models (Johnson-Laird, 1983) are a significant part of our understanding of language. Mental models suggest that the meanings learned by our agent should be based upon structured perception of the environment by the learner (Section 3.1) rather than clusters of raw sensor values.

To explore these ideas empirically, we have developed a phrase-learning system which uses the Multi-Stream Dependency Detection (MSDD) algorithm (Oates et al., 1999), (Section 4.1). MSDD ranks associations between tokens based on the G statistic, an information-theoretic criterion derivable from from mutual information. Using this learning system our mobile robot was able to learn a moderate-sized

vocabulary of phrases which describe a set of its actions (Section 5.).

2. Related Work

There has been a great deal of machine learning focused on natural language processing. Here we describe work that has focused on unsupervised acquisition of word meaning by an autonomous agent.

Oates (Oates et al., 1999) used mutual information to cluster words which have similar syntactic structure. Meaning is attributed to these clusters of words by estimating the probability that some cluster of words co-occurs with a sensory experience. Meanings are represented as clustered time series of the robot's raw sensors.

This work was later expanded by Oates (Oates, 2001) in the PERUSE algorithm which segmented raw speech data rather than textual descriptions. PERUSE finds patterns by running a window over the sound signal for each utterance and finding portions of this pattern in other utterances. It then fits a Gaussian temporal model to the patterns, learning the parameters using Expectation Maximization. Once words are identified, they are tagged. A separate algorithm estimates the conditional probability that the word would be uttered given a certain set of sensor values.

In Steels' (Steels, 1996) *talking heads experiment*, interacting agents evolve a language to accomplish a communication task. The agents randomly generated nonsense words and then negotiated their meaning with the other agent. The words in this language denoted specific areas on a scene which both agents were viewing. In contrast to any human language, none of the words in the language were context dependent. Each had a single meaning which was some descriptive attribute of the scene, for example `(size tall)`. Although "tall" is highly contextualized in English, in this work it was a specific category, programmatically defined. There was no model of the environment in the representation. The structure of the experiment, a game where the agents took turns telling the each other which element in a scene should be examined, was such that denotational meanings were sufficient to play the game and structured representations of the world were unnecessary.

In later work, (Steels and Kaplan, 2001), a robot plays a game with a human where it learns to associate the names of three objects with their perceptual representation. Associations are learned using reinforcement learning. The words are obtained from raw speech using off-the-shelf speech recognition software. The perceptions of the environment are clusters of simplified color histograms.

Deb Roy (Roy, 2000) built a system which found correspondences between utterances and visual input. The utterances were represented as phoneme probabilities matched with hidden markov models of phoneme transitions. Objects were represented as collections of histograms of both color and shape information. The agent was exposed to a series of distinct visual/auditory experiences. The agent maintained a queue of five of these experiences. Whenever at least two of the experiences in the queue were considered to be the same, the agent stored the utterance-visual pairing in its long term memory. At the end of the experiement the agent rated it's pairings by the mutual information between utterance and observation.

In all of this work, the meanings which are learned are denotational. Words are mapped to raw sensor data (via clusters, or other abstraction) for the purposes of recognizing objects/scenes. Partially this is a result of a focus in previous work on learning nouns, and the goal has been discrimination of the environment. Machine learning of linguistic constructs requiring richer representational structure, such as that described in the mental model literature, has not yet been addressed.

Jeff Siskind has done the work which is most closely related to the concept of functional meanings. His Abigail (Siskind, 1993) system learned to recocognize various event types described by predicate logic where the predicates were simple physical primitives. The system had a "imagination" which allowed it to make hypothesis about the future of the environment given its current condition. In contrast to our work, Abigail was a recognition system. Although it could simulate the world it could not behave pro-actively in it. Siskind's later work (Siskind, 2000) focused on building an algorithm which was an adequate explanation of children's word learning which addressed, among other things, the multi-word learning problem which is related to contextualized meanings for words.

3. Implications of the Structure of Meaning

Words contain Shannon information which can be harnessed to learn their meanings. The structure of the meanings we learn has a significant impact on the uses to which the meanings can be put, and, obviously, what meanings can be learned. These issues are addressed in the following.

In the last section, we discussed the denotational meanings used by previous associative word learners. However, a great deal of psychological evidence suggests that much of meaning, especially complex meaning, is based upon a mental model which goes far beyond a denotational token. Research shows that by the time they begin to use words, children represent the structure of events and processes, in addition to tracking objects (Bruner, 1983,

Nelson, 1986, Tomasello, 1992). Even the meaning of simple word like "cup" is not denotational. While the cluster of sensor inputs which indicates the class of appearances we might label "cup" is a part of the word's meaning, a far more important part of the meaning of "cup" is the knowledge that liquid, when poured into a "cup" will not run out all over our laps. Denotational meanings might allow a learner to say "cup" when one is presented to it, functional meanings allow the learner to drink (and more). Further, the learner's ability to model its meanings allows it to take a pro-active role in the learning process. It can say "I'm going to move forward now" and then do so, rather than simply waiting to be told that it has moved forward. Pro-active learning is closely related to the language games discussed in Steels and elsewhere.

3.1 Functional Meanings and Mental Models

Many psychologists have proposed that a large component of language comprehension is based upon mental models (Johnson-Laird, 1983, Gentner and Stevens, 1983). Speaking loosely, a mental model is a representation which contains what a word is "about" rather than the word itself. Denotational meanings do not consist of what the word is "about." Denotational meaning is simply a mapping between a word and some token. An especially clear example of this in previous work is Steels' talking heads whose learned meanings are the ability to perform a substitution of a perceptual token for a heard word.

There is, however, very little consensus as to the essential nature of the model itself. In our work, we have and intend to use models that seem necessary. We do not claim that they are complete, minimal, or even those actually present in humans. Further, it is our belief that many of these models are learned. Our purpose in implementing them for the learner is to facilitate more complex learning rather than to demonstrate the necessity of their presence *a priori* in a word learning system.

Functional meaning is grounded in control structures that allow the agent to act appropriately in the environment – e.g., with respect to an object or event. An example of such a control structure is a planning operator, which provides conditions for the appropriateness of an action to achieve some goal. When the functional representation of a phrase has been successfully learned, the agent can use the mapping of the word to action system or operator to either act appropriately or internally model the word's meaning.

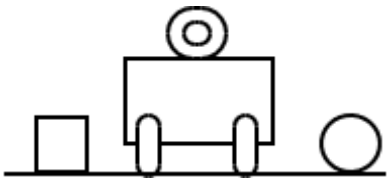

Figure 1: A hypothetical learning environment for a mobile robot

3.2 Learning Phrases Rather than Words

Previous work has focused upon learning the denotational meaning of individual words. However, learning meanings for individual words outside of context limits the concepts which can be learned.

Consider an agent trying to learn the meaning of three simple actions: turning left ninety degrees, turning right ninety degrees and staying still. Suppose further, that the agent's actions take place in an environment where to the left of the agent lies a sphere and to the right a cube. Figure 1 shows this hypothetical situation.

Suppose the robot's actions are chosen at random. When it chooses turn left it receives the description "The robot turned left." When it turns right it receives the description "The robot turned right." When it stays still, it receives the description "The robot sits still with a cube to the right and a sphere to the left." Given this data, and a bag of words representation of the utterances, an associative learner cannot learn the meaning of "left" or "right." There is no statistical correlation between turning left and the word "left" nor between the sphere present to the left. This occurs of course because the meanings are different depending on how it used in the sentence, e.g. "turned left" and "turned right" versus "to the left" and "to the right."

The lesson of this example is that if we are interested in capturing this meaning, we must consider words in context, not just words individually. Although the importance of syntactic context has long been recognized by linguists working in formal semantics, previous work in agent-based associative learning, as cited earlier, has not taken this into account. However such context must be considered for even simple understanding of acting in the world.

Of course, considering whole sentences makes the associative learning task quite difficult. One need only compare Hemingway and Faulkner to see that two sentences carrying essentially the same meaning may vary greatly in size and structure. To simplify the problem, but maintain our ability to learn contextualized meanings, we have chosen to encode our phrases as $(subject, verb, object)$ triples. This representation is analogous to a simplified case frame. The associations that we learn consist of a mapping

from tuple to experience. Each element of the tuple is either a specific word or the wildcard symbol '*' indicating that the meaning of the phrase remains constant for any substitution for the wildcard. Although this representation is greatly simplified in comparison to natural language sentences, it contains enough context to greatly expand the meanings we can learn.

4. Associative Word Learning through Mutual Information

In order to use the information which words convey about the environment to learn their meanings, we need an information-theoretic learning algorithm. We have chosen to use the Multi-Stream Dependency Detection algorithm (MSDD) developed by Oates et al (Oates et al., 1999). The MSDD algorithm uses the G statistic to learn associations. As we show below, the G statistic is directly related to the Shannon mutual information between two distributions.

4.1 MSDD

The MSDD algorithm (Oates et al., 1999) finds dependencies between vectors of tokens. The training data is a set of pairs $< \vec{x}, \vec{y} >$, where $\vec{x}$ and $\vec{y}$ are token vectors that need not have the same length. Often, MSDD has been used to find patterns in multivariate categorical time series, in which case the training pairs $\vec{x}$ and $\vec{y}$ would be samples from the time series separated by some lag.

MSDD outputs a set of a patterns. Each pattern has the form $< \vec{p}, \vec{q} >$, where each p_i, q_i can be either a token that occurred in the training data or the wildcard symbol $*$. A pattern vector $\vec{p}$ matches an input vector $\vec{q}$ when for all i, either $x_i = p_i$ or $p_i = *$. An MSDD pattern $< \vec{p}, \vec{q} >$ is interpreted as, for a pair $< \vec{x}, \vec{y} >$, if $\vec{x}$ matches $\vec{p}$, then $\vec{y}$ matches $\vec{q}$ more often than would be expected by chance.

MSDD evaluates candidate patterns using contingency tables. For a pattern $< \vec{p}, \vec{q} >$, let O(P,Q) be the number of pairs where $\vec{x}$ matches $\vec{p}$ and $\vec{y}$ matches $\vec{q}$. Let $O(P, \neg Q)$ be the number of pairs where $\vec{x}$ matches $\vec{p}$ and $\vec{y}$ does not match $\vec{q}$, T_1 be the statement that $\vec{x}$ matches $\vec{p}$, and let T_2 be the statement that $\vec{y}$ matches $\vec{q}$, then MSDD evaluates the pattern by constructing the contingency table shown in Figure 1.

	T_1	$\neg T_1$
T_2	$O(P,Q)$	$O(\neg P,Q)$
$\neg T_2$	$O(P,\neg Q)$	$O(\neg P,\neg Q)$

Table 1: An example contingency table used by the MSDD algorithm

If the G statistic for this contingency table is significant then we can dismiss the null hypothesis that T_1 and T_2 are independent and we can conclude that there is a statistical association between T_1 and T_2.

Obviously, the space of possible patterns is quite large. In order to obtain a solution in a reasonable amount of computation, the MSDD algorithm begins with T_1 and T_2 filled entirely with wildcards. It then proceeds with a guided search down the tree which is built by expanding patterns. Paterns are expanded by changing a wildcard into its possible values. This search tree is pruned by bounds on the G statistic.

4.2 Relating G to Mutual Information

The MSDD algorithm uses the G statistic to select word associations. We have noted that the G statistic is an information-theoretic measure. In the following we show that MSDD's ranking of associations by the G statistic is equivalent to ranking them by mutual information.

Suppose we have a sample drawn from the joint distribution of two random variables X and Y. Let $O(x, y)$ be the observed count for the pair (x, y), t be the total number of observations and $E(x, y)$ be the expected count under the hypothesis that X and Y are unrelated, that is:

$$E(x,y) = \frac{\left(\sum_x O(x,y)\right)\left(\sum_y O(x,y)\right)}{t}$$

Then G is defined as:

$$\begin{aligned} G &= 2\sum_x \sum_y O(x,y) \log \frac{O(x,y)}{\frac{\sum_y O(x,y) \times \sum_x O(x,y)}{t}} \\ &= 2t \sum_x \sum_y \frac{O(x,y)}{t} \log \frac{\frac{O(x,y)}{t}}{\frac{\sum_y O(x,y)}{t} \times \frac{\sum_x O(x,y)}{t}} \\ &\approx 2t \sum_x \sum_y P(x,y) \log \frac{P(x,y)}{P(x) \times P(y)} \\ &\approx 2tMI(x,y). \end{aligned}$$

So G is twice the sample size times the mutual information. As long as t is kept constant, maximizing G is equivalent to maximizing mutual information. The total number of observations is constant for any pair of tokens associated by MSDD, because there is a fixed number of pairs of observations in a data-set, and each of them fall into one of the four table cells for any pair of tokens. Therefore, any ranking of associations based upon the G statistic is equivalent to a ranking based upon mutual information.

5. Empirical Examination

5.1 Experimental Set-up

To test for evidence of Shannon information in words and to see how it might guide word learning, we ran an experiment on a Pioneer II mobile robot (Figure

2). The Pioneer II was given five primitive actions which it could perform: moving forward, moving backward, turning left ninety degrees, turning right ninety degrees and sitting still. Nine digital movies (two of each action except sitting still) were recorded. While the robot was engaged in each action, its perceptual system recorded the following perceptual vector *(heading-delta position-delta)*. *heading-delta* and *position-delta* were each one of the values: *negative, zero, positive*. These perceptions were not just denotational tokens. The agent could use a vector such as *(positive zero)* to perform a ninety-degree turn. A series of vectors such as *((positive zero) (zero positive) (negative zero) (zero positive))* constitute a plan which the agent could use to perform planned action. In this case a movement through a doorway and turning into a hall.

Each movie of the robot acting was shown to between eight and twelve people. Each person wrote a textual description of the movies, for example: "The robot rotates ninety degrees and stops, facing away from the viewer." Each of these textual descriptions were manually processed into *(subject, verb, object)* tuples. This resulted in eighty-one descriptions with a vocabulary of sixty different words; three words for subject (mostly "robot"), thirty-nine verbs and nineteen objects (including some tuples with a `nil` object).

MSDD was run on the data to find associations between phrases and perception vectors. We used MSDD with k-best tree pruning, where search proceeds breadth first through the space of associations until expanding a new level of the tree does not result in any changes to the k-best associations already found. Since all of the observations were actions, we also required that a verb be present in any of the associations which were learned.

5.2 Results

When MSDD was run to find the thirty best patterns it found the associations shown in Table 2. Twenty-seven are listed: Three other associations found are not shown because they are subsumed by others that are listed.

Afterward, the agent was given phrases such as "robot turns left" and "robot is moving forward" and was asked to replicate the correct behaviour indicated by the term. Since the agent had access to meanings which could be turned into actions (as described above), the entered phrase could easily be mapped to the agent's perceptual vector. With the perceptual vector in hand the agent could successfully produce the corresponding behaviour. Likewise since the agent maintained a short term memory of its past actions it could generate the phrases which corresponded to its previous actions.

Sensor Value (heading position)	Phrases Associated
(zero positive)	"* moves *", "* moving forward", "* moves forward", "* moving *"
(zero negative)	"* * backward", "* backed *", "* backs *"
(zero zero)	"* idles *", "* resting *", "* stays *", "* sleeps *", "* motionless *", "* standing *"
(positive zero)	"* turning right", "* turning *", "* turning clockwise", "* turns clockwise"
(negative zero)	"* turning left", "* spinning left", "* turns counter-clockwise", "* turns left", "* turns *"
(* zero)	"* turns *", "* turning *" "* rotates *"
(zero *)	"* moves *", "* moving *"

Table 2: Mapping of words associated with sensor values

5.3 Discussion

The first thing that is interesting to note is that the subject is always wildcarded in the learned phrases. This is unsurprising given that seventy-nine out of the eighty-one descriptions collected used "robot" as the subject. As a result, the subject was never closely correlated with any particular action. Obviously this is a byproduct of the robot being the subject in all of the actions. In the future we expect that exposure to situations where the subject varies will produce phrases whose meanings are distringuished by their subject.

Also of note are the rules for (* zero) and (zero *). In both cases these meanings indicate that the agent has the knowledge that (valid in this environment) turning means not moving forward and moving means not turning. However, these rules indicate a weakness in the MSDD algorithm's ability to learn associations. MSDD is forced to choose between a wildcard or a specific symbol. It would be preferable for the system to learn (non-zero zero) maps to "* turns *", where non-zero means a value of negative or positive. This type of expansion of MSDD's generalization abilities is an area that ought to be explored.

Although by and large the associations learned by MSDD are accurate meanings, three:
(* moves *) → (zero positive)
(* turning *) → (positive zero)
(* turns *) → (negative zero)
are incorrect. In each of these cases the meanings learned, while accurate, are overly specific. This is because the human annotators used "moves" much more often with "forward" than "backward". For moving backward, they were more likely to use

Figure 2: The robot engaged in the left turn action

"backs", "backing", etc. Likewise, "turns" was used with "left", while "turning" was used with right. These results are symptomatic of the very real problem that accidental correlation can (and most likely will) occur. This is an instance of over-fitting, a problem which plagues nearly all learning. In meaning acquisition, one potential solution is the addition of hypothesis testing by the learner, or a language game such as those proposed by Steels to provide a mediator to aid the learning system in correcting misinterpretations.

6. Conclusions

Information can be exploited by a learning algorithm to associate words with meanings. To learn meanings which are subsequently useful to the learning system, it is preferable to learn functional meanings rather than the denotational meanings which have been the focus of previous associative word learners. Functional meanings are also consistent with the theories of mental models developed in psychology. Functional meanings necessitate a learner with structured perception of the environment since acquired functional meanings must hang on a framework which can capture an action's potential effects on the world. More complex meanings also necessitate the association of phrases rather than words to meanings.

Initial experimental results show our agent is capable of learning functional meanings of phrases describing a subset of its actions.

6.1 Future Work

There are many ways we plan to expand this work. Initially we have dealt only with primitive actions on the part of the robot. We would like to expand this to include more complex planned actions. Additionally we see language as a tool that allows an agent to plan its actions. For example, a statement like "come here and pick this up" implicitly encodes a plan if you have a meaning for "come here" and "pick this up."

Like Steels, we believe that a language game is important for successful language acquisition, because they provide a tight feedback loop which directly correlates an agent's success or failure in the game to success or failure in language acquisition. Further the language game can necessitate the development of functional meanings, since a game can be designed which an agent using denotational meanings cannot play successfully.

We are working to increase the complexity of vocabulary and sentence grammar, as well as the complexity of activities. Our eventual goal is an account of full language development in a robotic platform. Along the way, we hope to identify the limits of associative language learning, seeing how such learning scales in the face of greater demands.

References

Bruner, J. S. (1983). *Child's talk: Learning to use language.* Norton.

Gentner, D. and Stevens, A., (Eds.) (1983). *Mental Models.* Erlbaum.

Johnson-Laird, P. (1983). *Mental Models.* Harvard University Press.

Nelson, K. (1986). *Event Knowledge: Structure and function in development.* Academic.

Oates, T. (2001). *Grounding Knowledge in Sensors: Unsupervised Learning for Language and Planning.* PhD thesis, University of Massachusetts, Amherst.

Oates, T., Eyeler-Walker, Z., and Cohen, P. (1999). Using syntax to learn semantics: An experiment in language acquisition with a mobile robot. Technical report, Department of Computer Science, University of Massachusetts.

Roy, D. (2000). Integration of speech and vision using mutual information. In *International Conference on Acoustics, Speech and Signal Processing.*

Siskind, J. (1993). Grounding language in perception. In *Proceedings of SPIE.*

Siskind, J. (2000). *Models of Language Acquisition: Inductive and Deductive Approaches*, chapter Learning Word-to-Meaning Mappings. Oxford University Press.

Steels, L. (1996). Perceptually grounded meaning creation. In *International Conference on Multi-Agent Systems*, pages 562–567.

Steels, L. and Kaplan, F. (2001). Aibo's first words, the social learning of language and meaning. *Evolution of Communication*, 4(1).

Tomasello, M. (1992). *First Verbs: A case study of early grammatical development.* Cambridge University Press.