Grounding Word Meanings in Sensor Data: Dealing with Referential Uncertainty

Tim Oates
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
1000 Hilltop Circle
Baltimore, MD 21250
oates@cs.umbc.edu

Abstract

We consider the problem of how the meanings of words can be grounded in sensor data. A probabilistic representation for the meanings of words is defined, a method for recovering meanings from observational information about word use in the face of referential uncertainty is described, and empirical results with real utterances and robot sensor data are presented.

1 Introduction

We are interested in how robots might learn language given qualitatively the same inputs available to children - natural language utterances paired with sensory access to the environment. This paper focuses on the sub-problem of learning word meanings. Suppose a robot has acquired a set of sound patterns that may or may not correspond to words. How is it possible to separate the words from the non-words, and to learn the meanings of the words?

We assume the robot's sensory access to its environment is through a collection of primitive sensors organized into sensor groups, where each sensor group is a set of related sensors. For example, the sensor group Ψ_G might return a single value representing the mean grayscale intensity of a set of pixels corresponding to an object in the visual field. The sensor group Ψ_{HW} might return two values representing the height and width of the bounding box around the object.

Learning the meanings of words requires a representation for meaning. We use a representation that we call a conditional probability field (CPF), which is a type of scalar field. A scalar field is a map of the following form:

$$f : \Re^n \mapsto \Re$$

The mapping assigns to each vector $\mathbf{x} \in \Re^n$ a scalar value $f(\mathbf{x})$. A conditional probability field assigns to each $\mathbf{x}$, which corresponds to a point in an n-dimensional sensor group, a conditional probability of the form $p(\mathtt{E}|\mathbf{x})$, where $\mathtt{E}$ denotes the occurrence of some event. Let $\mathtt{E}(\Psi, \mathbf{x})$ denote the CPF defined over sensor group Ψ for event $\mathtt{E}$.

The semantics of a CPF clearly depend on the nature of $\mathtt{E}$. Two events that will be of particular importance in learning the meanings of words are:

- `utter-W` - the event that word W is uttered, perhaps as part of an utterance that refers to some feature of the world denoted by W
- `hear-W` - the event that word W is heard

The corresponding conditional probability fields are:

- `utter-W`$(\Psi, \mathbf{x})$ - the probability that word W will be uttered by a competent speaker of the language to denote the feature of the physical world that Ψ is currently sensing (i.e. that results in the current value of $\mathbf{x}$)
- `hear-W`$(\Psi, \mathbf{x})$ - the probability that word W will be heard given that $\mathbf{x} \in \Psi$ is observed

In this framework, the meaning of word W is simply `utter-W`$(\Psi, \mathbf{x})$. The last plot in figure 3 shows a CPF defined over Ψ_G that might represent the meaning of the word "gray". Grayscale intensities near 128 will be called gray with probability almost one, whereas intensities near 0 and 255 will never be called gray. Rather, they are "black" and "white" respectively.

Learning the denotation of W involves determining the identity of Ψ and then recovering `utter-W`$(\Psi, \mathbf{x})$. The learner does not have direct access to `utter-W`$(\Psi, \mathbf{x})$. Rather, the learner must gain information about `utter-W`$(\Psi, \mathbf{x})$ indirectly, by noticing the sensory contexts in which W is used and those in which it is not, i.e. via `hear-W`$(\Psi, \mathbf{x})$.

This problem is difficult due to referential uncertainty. Even if the utterances the learner hears are true statements about aspects of its environment that are perceptually available, there are are usually many aspects of the environment that might be a given word's referent. This is Quine's "gavagai" problem (Quine, 1960). The algorithm described in this paper solves a restricted version of the gavagai problem, one in which the denotation of a word must be representable as a CPF defined over one of a set of pre-defined sensor groups.

2 A Simplified Learning Problem

Rather than starting with the full complexity of the problem facing the learner, consider the following highly simplified version. Suppose an agent with a single sensor group, Ψ_G, lives in a world with a single object, o_1, that periodically changes color. Each time the color changes, a one word utterance is generated describing the new color, which is one of "black", "white" or "gray".

In this scenario there is no need to identify Ψ because there is only one possibility. Also, each time a word is uttered there is perfect information about its denotation; it is the current value produced by $\Psi_G(o_1)$. (The notation $\Psi(o)$ indicates the value recorded by sensor group Ψ when it is applied to object o. This assumes an ability to individuate objects in the environment.) Therefore, the probability that a native speaker of our simple language will utter W to refer to $\mathbf{x}$ is the same as the probability of hearing W given $\mathbf{x}$. This fact makes it possible to recover the form of the CPF for each of the three words by noticing which values of $\Psi_G(o_1)$ co-occur with the words and applying Bayes' rule as follows:

$$\begin{aligned} \texttt{utter-W}(\Psi_G, \mathbf{x}) &= \texttt{hear-W}(\Psi_G, \mathbf{x}) \\ &= p(\texttt{hear-W}|\mathbf{x}) \\ &= \frac{p(\mathbf{x}|\texttt{hear-W})p(\texttt{hear-W})}{p(\mathbf{x})} \end{aligned}$$

The maximum-likelihood estimate of the quantity $p(\texttt{hear-W})$ is simply the number of utterances containing W divided by the total number of utterances. The quantities $p(\mathbf{x})$ and $p(\mathbf{x}|\texttt{hear-W})$ can be estimated using a number of standard techniques. We use kernel density estimators with (multivariate) Gaussian kernels to estimate probability densities such as these.

The simplified version of the word-learning problem presented in this section can be made more realistic, and thus more complex, by increasing either the number of objects in the environment or the number of sensor groups available to the agent. Section 3 explores the former, and section 4 explores the latter.

3 Multiple Objects

When there is no ambiguity about the referent of a word, it is possible to recover the conditional probability field that represents the word's denotational meaning by passive observation of the contexts in which it is used. Unfortunately, referential ambiguity is a feature of natural languages that we contend with on a daily basis. This ambiguity appears to be at its most extreme for young children acquiring their first language who must determine for each newly identified word the referent of the word from the infinitely many aspects of their environment that are perceptually available.

Consider what happens when we add a second object, o_2, to our example domain. If both objects change color at exactly the same time, though not necessarily to the same color, the learner has no way of knowing whether an utterance refers to the value produced by $\Psi_G(o_1)$ or $\Psi_G(o_2)$. In the absence of any exogenous information about the referent, the best the learner can do is make a guess, which will be right only 50% of the time. As the number of objects in the environment increases, this percentage decreases.

Referential ambiguity can also take the form of uncertainty about the sensor group to which a word refers. Given two objects, o_1 and o_2, and two sensor groups, Ψ_1 and Ψ_2, a word can refer to any of the following: $\Psi_1(o_1)$, $\Psi_1(o_2)$, $\Psi_2(o_1)$, $\Psi_2(o_2)$. In this section we make the unrealistic assumption that the learner has a priori knowledge about the sensor group to which a word refers. This assumption will be relaxed in the following section.

Intuitively, referential ambiguity clouds the relationship between the denotational meaning of a word and the observable manifestations of its meaning, i.e. the contexts in which the word is used. As we will now demonstrate, it is possible to make the nature of this clouding precise, leading to an understanding of the impact of referential ambiguity on learnability.

Suppose an agent hears an utterance U containing word W while its attention is focused on the output of sensor group Ψ. (Recall that in this section we are making the assumption that the agent knows that W refers to Ψ.) Why might U contain W? There are two mutually exclusive and exhaustive cases: U is (at least in part) about Ψ, and W is chosen to denote the current value produced by Ψ; U is not about Ψ, and U contains W despite this fact. The latter case might occur if, for example, W has multiple meanings and the utterance uses one of the meanings of W that does not denote a value produced by Ψ.

Let $\texttt{about}(U, \Psi)$ denote the fact that U is (at least in part) about Ψ, and let $\texttt{contains}(U, W)$ denote the fact that W occurs in U. Then the conditional probability of an utterance containing W given the current value, $\mathbf{x}$,

$$
\begin{aligned}
\texttt{hear-W}(\Psi,\mathbf{x}) &= p(\texttt{about}(U,\Psi) \wedge \texttt{contains}(U,W) \vee \neg\texttt{about}(U,\Psi) \wedge \texttt{contains}(U,W)) && (1)\\
&= p(\texttt{about}(U,\Psi) \wedge \texttt{contains}(U,W)) + p(\neg\texttt{about}(U,\Psi) \wedge \texttt{contains}(U,W)) - \\
&\quad p(\texttt{about}(U,\Psi) \wedge \texttt{contains}(U,W) \wedge \neg\texttt{about}(U,\Psi) \wedge \texttt{contains}(U,W)) \\
&= p(\texttt{about}(U,\Psi) \wedge \texttt{contains}(U,W)) + p(\neg\texttt{about}(U,\Psi) \wedge \texttt{contains}(U,W)) \\
&= p(\texttt{contains}(U,W)|\texttt{about}(U,\Psi))p(\texttt{about}(U,\Psi)) + \\
&\quad p(\texttt{contains}(U,W)|\neg\texttt{about}(U,\Psi))p(\neg\texttt{about}(U,\Psi)) \\
&= \alpha\texttt{utter-W}(\Psi,\mathbf{x}) + (1-\alpha)\beta && (2)
\end{aligned}
$$

Figure 1: A derivation of the relationship between $\texttt{hear-W}(\Psi,\mathbf{x})$ and $\texttt{utter-W}(\Psi,\mathbf{x})$.

produced by Ψ can be expressed via equation 1 in figure 1. Equation 1 is a more formal, probabilistic statement of the conditions given above under which U will contain W. It can be simplified as shown in the remainder of the figure.

The first step in transforming equation 1 into equation 2 is to apply the fact that $p(A \vee B) = p(A) + p(B) - p(A \wedge B)$. The resulting joint probability is the probability of a conjunction of terms that contains both $\texttt{about}(U,\Psi)$ and $\neg\texttt{about}(U,\Psi)$, and is therefore 0 and can be dropped. The remaining two terms are then rewritten using Bayes' rule. Finally, three substitutions are made:

- $\texttt{utter-W}(\Psi,\mathbf{x}) = p(\texttt{contains}(U,W)|\texttt{about}(U,\Psi))$
- $\alpha = p(\texttt{about}(U,\Psi))$
- $\beta = p(\texttt{contains}(U,W)|\neg\texttt{about}(U,\Psi))$

Simplification then leads directly to equation 2.

Before discussing the implications of equation 2, consider the import of α and β. The probability that U is about Ψ (i.e. α) is the probability that the speaker and the hearer are attending to the same sensory information. When $\alpha = 1$, there is perfect shared attention, and the speaker always refers to those aspects of the physical environment to which the hearer is currently attending. When $\alpha = 0$, there is never shared attention, and the speaker always refers to aspects of the environment other than those to which the hearer is currently attending.

The probability that U contains W even when U is not about Ψ (i.e. β) is the probability that W will be used to refer to some feature of the environment other than that measured by Ψ. There are two reasons why W might occur in a sentence that does not refer to Ψ:

- W is polysemous and one of the meanings that does not refer to Ψ is used in the utterance
- W is used to refer to the value produced by Ψ for some object other than the one that is the hearer's focus of attention (e.g. $\Psi(o_1)$ rather than $\Psi(o_2)$)

Note that β comes into play only when $\alpha < 1$, i.e. when there is less than perfect shared attention between the speaker and the hearer.

The most significant aspect of equation 2 is from the standpoint of learnability. In our original one-object, one-sensor world there was never any doubt as to the referent of a word, and it was therefore the case that $\texttt{utter-W}(\Psi,\mathbf{x}) = \texttt{hear-W}(\Psi,\mathbf{x})$. This equivalence becomes clear in equation 2 by setting $\alpha = 1$ and simplifying. Because it is possible to compute $\texttt{hear-W}(\Psi,\mathbf{x})$ from observable information via Bayes' rule, it was possible in that world to recover $\texttt{utter-W}(\Psi,\mathbf{x})$ rather directly. However, equation 2 tells us that even in the face of imperfect shared attention (i.e. $\alpha < 1$) and homonymy (i.e. $\beta > 0$) it is the case that $\texttt{hear-W}(\Psi,\mathbf{x})$ is a linear transform of $\texttt{utter-W}(\Psi,\mathbf{x})$. Moreover, the values of α and β determine the precise nature of the transform.

To get a better handle on the effects of α and β on the manifestation of $\texttt{utter-W}(\Psi,\mathbf{x})$ through $\texttt{hear-W}(\Psi,\mathbf{x})$, consider figures 2 and 3. The last plot in figure 2 shows an example of a conditional probability field $\texttt{utter-W}(\Psi,\mathbf{x})$, which is also a plot of $\texttt{hear-W}(\Psi,\mathbf{x})$ when $\alpha = 1$. Figures 2 and 3 demonstrate the effects of varying α and β on $\texttt{hear-W}(\Psi,\mathbf{x})$. That is, the figures show how varying α and β affect the information about $\texttt{utter-W}(\Psi,\mathbf{x})$ available to the learner.

Recall from equation 2 that the conditional probability that word W will be heard given the current value produced by Ψ is a linear function of $\texttt{utter-W}(\Psi,\mathbf{x})$ which has slope α and intercept $(1-\alpha)\beta$. When the slope is zero (i.e. $\alpha = 0$) the speaker and the hearer never focus on the same features of the environment, and the probability of hearing W is just the background probability of hearing W, independent of the value of Ψ. When the slope is one (i.e. $\alpha = 1$) the speaker and the hearer always focus on the same features of the environment and so the effect of β vanishes. The observable manifestation of $\texttt{utter-W}(\Psi,\mathbf{x})$ and $\texttt{utter-W}(\Psi,\mathbf{x})$ are equivalent. These two case are shown in the first and last graphs in

figure 2, which contains plots of $\texttt{hear-W}(\Psi, \mathbf{x})$ over a range of values of β for various fixed values of α.

Figure 2 makes it clear that decreasing α preserves the overall shape of $\texttt{utter-W}(\Psi, \mathbf{x})$ as observed through $\texttt{hear-W}(\Psi, \mathbf{x})$, while squashing it to fit in a smaller range of values. Increasing α diminishes the effect of β, which is to offset the entire curve vertically. That is, the higher the level of shared attention between speaker and hearer, the less the impact of the background frequency of W on the observable manifestation of $\texttt{utter-W}(\Psi, \mathbf{x})$.

Figure 3, which shows plots of $\texttt{hear-W}(\Psi, \mathbf{x})$ given a range of values of α for various fixed values of β, is another way of looking at the same data. The role of α in squashing the observable manifestation of $\texttt{utter-W}(\Psi, \mathbf{x})$ is apparent, as is the role of β in vertically shifting the curves. Only when $\alpha = 0$ is there no information about the form of $\texttt{utter-W}(\Psi, \mathbf{x})$ in the plot of $\texttt{hear-W}(\Psi, \mathbf{x})$.

What does all of this have to say about the impact of α and β on the learnability of word meanings from sensory information about the contexts in which they are uttered? As we will demonstrate shortly, if the following expression is true for a given conditional probability field, $\texttt{utter-W}(\Psi, \mathbf{x})$, then it is possible to recover that CPF from observable data (i.e. from $\texttt{hear-W}(\Psi, \mathbf{x})$):

$$\exists \mathbf{x_0}\ \texttt{utter-W}(\Psi, \mathbf{x_0}) = 0 \wedge \exists \mathbf{x_1}\ \texttt{utter-W}(\Psi, \mathbf{x_1}) = 1$$

The claim is as follows. If there is both a value produced by Ψ that is always referred to as W and a value that is never referred to as W, one can recover the CPF that represents the denotational meaning of W simply by observing the contexts in which W is used.

Intuitively, the above expression places two constraints on word meanings. First, for a word W whose denotation is defined over sensor group Ψ, it must be the case that some value produced by Ψ is (almost) universally agreed to have no better descriptor than W; there is no other word in the language that is more suitable for denoting this value. Second, there must be some value produced by Ψ for which it is (almost) universally agreed that W is not the best descriptor. It is not necessarily the case that W is the worst descriptor, only that some other word or words are better.

As equation 2 indicates, $\texttt{hear-W}(\Psi, \mathbf{x})$ is a linear transform of $\texttt{utter-W}(\Psi, \mathbf{x})$ with slope α and intercept $(1 - \alpha)\beta$. If we know two points on the line defined by equation 2 we can determine its parameters, making it possible to reverse the transform and compute the value of $\texttt{utter-W}(\Psi, \mathbf{x})$ given the value of $\texttt{hear-W}(\Psi, \mathbf{x})$.

Because $\texttt{hear-W}(\Psi, \mathbf{x})$ is a linear transform of $\texttt{utter-W}(\Psi, \mathbf{x})$, any value of $\mathbf{x}$ that minimizes (maximizes) one minimizes (maximizes) the other. Recall that conditional probability fields map from sensor vectors to probabilities, which must lie in the range [0, 1]. Under the assumption that $\texttt{utter-W}(\Psi, \mathbf{x})$ takes on the value 0 at some point, such as when $\mathbf{x} = \mathbf{x}_0$, $\texttt{hear-W}(\Psi, \mathbf{x})$ must be at its minimum value at that point as well. Let that value be p_{min}. Likewise, under the assumption that $\texttt{utter-W}(\Psi, \mathbf{x})$ takes on the value 1 at some point, such as when $\mathbf{x} = \mathbf{x}_1$, $\texttt{hear-W}(\Psi, \mathbf{x})$ must be at its maximum value at that point as well. Let that value be p_{max}. These observations lead to the following system of two equations:

$$\begin{aligned} p_{min} &= \alpha * \texttt{utter-W}(\Psi, \mathbf{x_0}) + (1 - \alpha)\beta \\ &= \alpha * 0 + (1 - \alpha)\beta \\ &= (1 - \alpha)\beta \\ p_{max} &= \alpha * \texttt{utter-W}(\Psi, \mathbf{x_1}) + (1 - \alpha)\beta \\ &= \alpha * 1 + (1 - \alpha)\beta \\ &= \alpha + (1 - \alpha)\beta \end{aligned}$$

Solving these equations for α and β yields the following:

$$\begin{aligned} \alpha &= p_{max} - p_{min} \\ \beta &= \frac{p_{min}}{1 - p_{max} + p_{min}} \end{aligned}$$

Recall that the goal of this exercise is to recover $\texttt{utter-W}(\Psi, \mathbf{x})$ from its observable manifestation, $\texttt{hear-W}(\Psi, \mathbf{x})$. This can finally be accomplished by substituting the values for α and β given above into equation 2 and solving for $\texttt{utter-W}(\Psi, \mathbf{x})$ as shown in figure 4.

That is, one can recover the CPF that represents the denotational meaning of a word by simply scaling the range of conditional probabilities of the word given observations so that it completely spans the interval $[0, 1]$.

4 Multiple Sensor Groups

This section considers a still more complex version of the problem by allowing the learner to have more than one sensor group. Suppose an agent has two sensor groups, Ψ_1 and Ψ_2, and that word W refers to Ψ_1. The agent can observe the values produced by both sensor groups, note whether each value co-occurred with an utterance containing W, compute both $\texttt{hear-W}(\Psi_1, \mathbf{x})$ and $\texttt{hear-W}(\Psi_2, \mathbf{x})$, and apply equation 2 to obtain $\texttt{utter-W}(\Psi_1, \mathbf{x})$ and $\texttt{utter-W}(\Psi_2, \mathbf{x})$.

How is the agent to determine that $\texttt{utter-W}(\Psi_1, \mathbf{x})$ represents the meaning of W and $\texttt{utter-W}(\Psi_2, \mathbf{x})$ is garbage? The key insight is that if the meaning of W is grounded in Ψ, there will be some values of $\mathbf{x} \in \Psi$ for which it is more likely that W will be uttered than for others, and thus there will be some values for which it is more likely that W will be heard than others. Indeed, our ability to recover $\texttt{utter-W}(\Psi, \mathbf{x})$ from $\texttt{hear-W}(\Psi, \mathbf{x})$ is founded on the assumption that there is some value of $\mathbf{x} \in \Psi$ for which the conditional probability of uttering

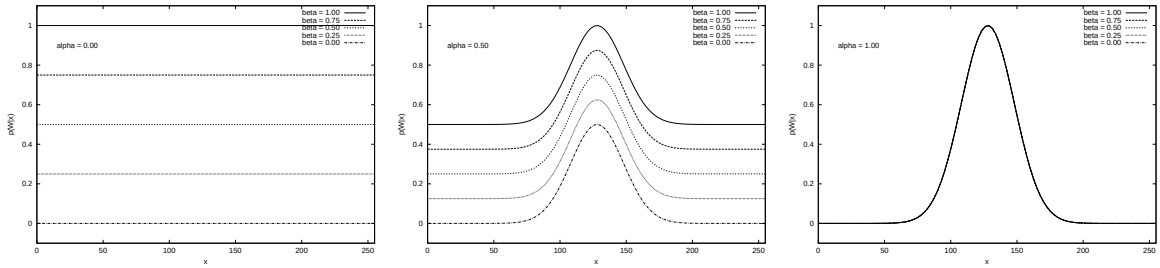

Figure 2: The effects of β on $\texttt{hear-W}(\Psi, \mathbf{x})$ for various values of α.

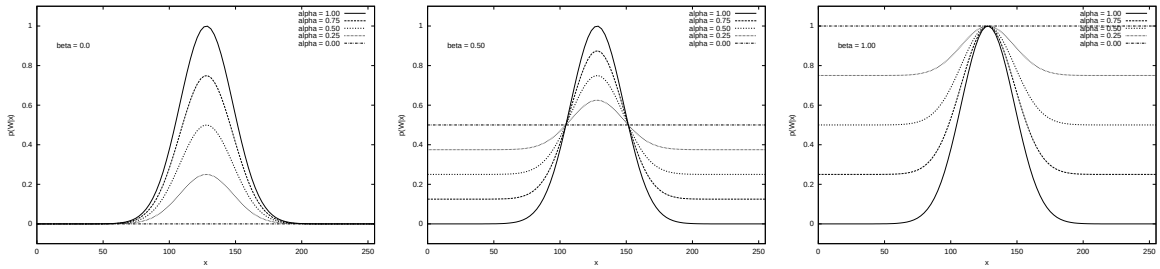

Figure 3: The effects of α on $\texttt{hear-W}(\Psi, \mathbf{x})$ for various values of β.

$$
\begin{aligned}
\texttt{hear-W}(\Psi, \mathbf{x}) &= \alpha\texttt{utter-W}(\Psi, \mathbf{x}) + (1-\alpha)\beta \\
&= (p_{max} - p_{min})\texttt{utter-W}(\Psi, \mathbf{x}) + (1 - p_{max} + p_{min})\frac{p_{min}}{1 - p_{max} + p_{min}} \\
&= (p_{max} - p_{min})\texttt{utter-W}(\Psi, \mathbf{x}) + p_{min} \\
\texttt{utter-W}(\Psi, \mathbf{x}) &= \frac{\texttt{hear-W}(\Psi, \mathbf{x}) - p_{min}}{p_{max} - p_{min}}
\end{aligned}
\tag{3}
$$

Figure 4: How to derive the meaning of a word from observations of its use in the face of referential uncertainty.

W is zero and some other value for which that probability is one. This is not necessarily the case for the conditional probability of hearing W given $\mathbf{x} \in \Psi$ because the level of shared attention between the speaker and the learner, α, influences the range of probabilities spanned by $\texttt{hear-W}(\Psi, \mathbf{x})$, with smaller values of α leading to smaller ranges.

Note that in our simple example with two sensor groups the speaker considers only the value of Ψ_1 when determining whether to utter W, and the learner considers only the value of Ψ_2 when constructing $\texttt{hear-W}(\Psi_2, \mathbf{x})$. Using the terminology and notation developed in section 3, there is no shared attention between the speaker and the learner with respect to W and Ψ_2, and it is therefore the case that $\alpha = 0$ and $\texttt{hear-W}(\Psi_2, \mathbf{x}) = \beta$.

If the exact value of $\texttt{hear-W}(\Psi_2, \mathbf{x})$ is known for all $\mathbf{x}$, an obviously unrealistic assumption, it is a simple matter to determine that $\texttt{utter-W}(\Psi_2, \mathbf{x})$ cannot represent the meaning of W by noting that it is constant. If $\texttt{utter-W}(\Psi_2, \mathbf{x})$ is not constant, then the speaker is more likely to utter W for some values of $\mathbf{x} \in \Psi_2$ than for others, and the meaning of W is therefore grounded in Ψ_2. As indicated by figure 2, the height of the bumps in the conditional probability field depend on α, the level of shared attention, but if there are any bumps at all we know that the meaning of W is grounded in the corresponding sensor group and we can recover the underlying conditional probability field. Under the assumption that the exact value of $\texttt{hear-W}(\Psi, \mathbf{x})$ can be computed, an agent can identify the sensor group in which the denotation of a word is grounded by simply recovering $\texttt{utter-W}(\Psi, \mathbf{x})$ for each of its sensor groups and looking for the one that is not constant.

In practice, the exact value of $\texttt{hear-W}(\Psi, \mathbf{x})$ will not be known, and it must be estimated from a finite number of observations. That is, an estimate of $\texttt{hear-W}(\Psi, \mathbf{x})$ will be used to compute an estimate of $\texttt{utter-W}(\Psi, \mathbf{x})$. Even if there is no association between W and Ψ, and $\texttt{utter-W}(\Psi, \mathbf{x})$ is therefore truly constant, an estimate of this conditional probability based on finitely many data will invariably not be constant. Therefore, the strategy of identifying relevant sensor groups by looking for bumpy conditional probability fields will not work.

The problem is that for any given word W and sensor group Ψ, it is difficult to distinguish between cases in which W and Ψ are unrelated and cases in which the meaning of W is grounded in Ψ but shared attention is low. The solution to this problem has two parts, both of which will be described in detail shortly. First, the mutual information between occurrences of words and sensor values will be used as a measure of the degree to which hearing W depends on the value produced by Ψ, and vice versa. Second, a non-parametric statistical test based on randomization testing will be used to convert the real-valued mutual information into a binary decision as to whether or not the denotation of W is grounded in Ψ.

4.1 Mutual Information

Let $I(W; \Psi)$ denote the mutual information between occurrences of word W and values produced by sensor group Ψ. The value of $I(W; \Psi)$ is defined as follows:

$$\int_{\mathbf{x} \in \Psi} p(\texttt{hear-W}, \mathbf{x}) \log \frac{p(\texttt{hear-W}, \mathbf{x})}{p(\texttt{hear-W})p(\mathbf{x})} dx + \int_{\mathbf{x} \in \Psi} p(\neg\texttt{hear-W}, \mathbf{x}) \log \frac{p(\neg\texttt{hear-W}, \mathbf{x})}{p(\neg\texttt{hear-W})p(\mathbf{x})} dx$$

Note that $I(W; \Psi)$ is the mutual information between two different types of random variables, one discrete (W) and one continuous (Ψ). In the expression above, the summation over the two possible values of W, i.e. $\texttt{hear-W}$ and $\neg\texttt{hear-W}$, is unpacked, yielding a sum of two integrals over the values of $\mathbf{x} \in \Psi$. Within each integral the value of W is held constant. Finally, recall that $\mathbf{x}$ is a vector with the same dimensionality as the sensor group from which it is drawn, so the integrals above are actually defined to range over all of the dimensions of the sensor group.

When $I(W, \Psi)$ is zero, knowing whether W is uttered provides no information about the value produced by Ψ, and vice versa. When $I(W, \Psi)$ is large, knowing the value of one random variable leads to a large reduction in uncertainty about the value of the other. Larger values of mutual information reflect tighter concentrations of the mass of the joint probability distribution and thus higher certainty on the part of the agent about both the circumstances in which it is appropriate to utter W and the denotation of W when it is uttered.

Although mutual information provides a measure of the degree to which W and Ψ are dependent, to understand and generate utterances containing W the agent must at some point make a decision as to whether or not its meaning is in fact grounded in Ψ. How is the agent to make this determination based on a single scalar value? The next section describes a way of converting scalar mutual information values into binary decisions as to whether a word's meaning is grounded in a sensor group that avoids all of the potential pitfalls just described.

4.2 Randomization Testing

Given word W, sensor group Ψ, and their mutual information $I(W; \Psi)$, the task facing the learner is to determine whether the meaning of W is grounded in Ψ. This can be phrased as a yes-or-no question in the following two ways. Is it the case that occurrences of W and the values produced by Ψ are dependent? Is it the case that

occurrences of W and the values produced by Ψ are not independent?

The latter question is the form used in statistical hypothesis testing. In this case the null hypothesis, H_0, would be that occurrences of W and the values produced by Ψ are independent. Given a distribution of mutual information values derived under H_0, it is possible to determine the probability of getting a mutual information value at least as large as $I(W;\Psi)$. If this probability is small, then the null hypothesis can be rejected with a correspondingly small probability of making an error in doing so (i.e. the probability of committing a type-I error is small). That is, the learner can determine that occurrences of W and the values produced by Ψ are not independent, that the meaning of W is grounded in Ψ, with a bounded probability of being wrong.

We've now reduced the problem to that of obtaining a distribution of values of $I(W;\Psi)$ under H_0. For most exotic distributions, such as this one, there is no parametric form. However, in such cases it is often possible to obtain an empirical distribution via a technique know as randomization testing (Cohen, 1995; Edgington, 1995).

This approach can be applied to the current problem as follows - each datum corresponds to an utterance and indicates whether W occurred in the utterance and the value produced by Ψ at the time of the utterance; the test statistic is $I(W;\Psi)$; and the null hypothesis is that occurrences of W and values produced by Ψ are independent. If the null hypothesis is true, then whether or not a particular value produced by Ψ co-occurred with W is strictly a matter of random chance. It is therefore a simple matter to enforce the null hypothesis by splitting the data into two lists, one containing each of the observed sensor values and one containing each of the labels that indicates whether or not W occurred, and creating a new data set by repeatedly randomly selecting one item from each list without replacement and pairing them together. This gives us all of the elements required by the generic randomization testing procedure described above.

Given a word and a set of sensor groups, randomization testing can be applied independently to each group to determine whether it is the one in which the meaning of W is grounded. The answer may be in the affirmative for zero, one or more sensor groups. None of these outcomes is necessarily right or wrong. As noted previously, it may be that the meaning of the word is too abstract to ground out directly in sensory data. It may also be the case that a word has multiple meanings, each of which is grounded in a different sensor group, or a single meaning that is grounded in multiple sensor groups.

5 Experiments

This section presents the results of experiments in which word meanings are grounded in the sensor data of a mobile robot. The domain of discourse was a set of blocks. There were 32 individual blocks with one block for each possible combination of two sizes (small and large), four colors (red, blue, green and yellow) and four shapes (cone, cube, sphere and rectangle).

To generate sensor data for the robot, one set of human subjects played with the blocks, repeatedly selecting a subset of the blocks and placing them in some configuration in the robot's visual field. The only restrictions placed on this activity were that there could be no more than three blocks visible at one time, two blocks of the same color could not touch, and occlusion from the perspective of the robot was not allowed.

Given a configuration of blocks, the robot generated a digital image of the configuration using a color CCD camera and identified objects in the image as contiguous regions of uniform color. Given a set of objects, i.e. a set of regions of uniform color in the robot's visual field, virtual sensor groups implemented in software extracted the following information about each object: Ψ_A measured the area of the object in pixels; Ψ_{HW} measured the height and width of the bounding box around the object; Ψ_{XY} measured the X and Y coordinates of the centroid of the object in the visual field; Ψ_{HSI} measured the hue, saturation and intensity values averaged over all pixels comprising the object; Ψ_S returned a vector of three numbers that represented the shape of the object (Stollnitz et al., 1996). In addition, the Ψ_{PCD} sensor group returned the proximal orientation, center of mass orientation and distance for the pair of objects as described in (Regier, 1996). These sensor groups constitute the entirety of the robot's sensorimotor experience of the configurations of blocks created by the human subjects.

From the 120 block configurations created by the four subjects, a random sample of 50 of configurations was shown to a different set of subjects who were asked to generate natural language utterances describing what they saw. The only restriction placed on the utterances was that they had to be truthful statements about the scenes.

Recurring patterns were discovered in the audio waveforms corresponding to the utterances (Oates, 2001) and these patterns were used as candidate words. Recall that a sensor group is semantically associated with a word when the mutual information between occurrences of the word and values in the sensor groups are statistically significant. Table 1 shows the p values for the mutual information for a number of combinations of words and sensor groups. Note from the first column that it is clear that the meaning of the word "red" is grounded in the Ψ_{HSI} sensor group. It is the only one with a statistically significant mutual information value. As the second column indicates, the mutual information between the word "small" and the Ψ_A sensor group is significant at the 0.05 level,

Table 1: For each sensor group and several words, the cells of the table show the probability of making an error in rejecting the null hypothesis that occurrences of the word and values in the sensor group are independent.

Sensor Group	Word		
	"red"	"small"	"above"
Ψ_A	0.76	0.05	0.47
Ψ_{HW}	0.86	0.09	0.31
Ψ_{XY}	0.29	0.67	0.07
Ψ_{HSI}	0.00	0.49	0.82
Ψ_S	0.34	0.58	0.44
Ψ_{PCD}	0.57	0.97	0.00

and the mutual information between this word and the Ψ_{HW} sensor group is not significant but is rather small. Both of these sensor groups return information about the size of an object, but the Ψ_{HW} sensor group overestimates the area of non-rectangular objects because it returns the height and width of a bounding box around an object. Finally, note from the third column that the denotation of the word "above" is correctly determined to lie in the Ψ_{PCD} sensor group, yet there appears to be some relationship between this word and the Ψ_{XY} sensor group. The reason for this is that objects that are said to be "above" tend to be much higher in the robot's visual field than all of the other objects.

How is it possible to determine the extent to which a machine has discovered and represented the semantics of a set of words? We are trying to capture semantic distinctions made by humans in natural language communication, so it makes sense to ask a human how successful the system has been. This was accomplished as follows. For each word for which a semantic association was discovered, each of the training utterances that used the word were identified. For the scene associated with each utterance, the CPF underlying the word was used to identify the most probable referent of the word. For example, if the word in question was "red", then the mean HSI values of all objects in the scene would be computed and the object for which the underlying CPF defined over HSI space yielded the highest probability would be deemed to be the referent of that word in that scene. A human subject was then asked if it made sense for that word to refer to that object in that scene.

The percentage of content words (i.e. words like "red" and "large" as opposed to "oh" and "there") for which a semantic association was discovered was 83.3%. Given a semantic association, the two ways that it can be in error are as follows: either the wrong sensor group is selected or the conditional probability field defined over that sensor group is wrong. Given all of the configurations for which a particular word was used, the semantic accuracy is the percentage of configurations that the meaning component of the word selects an aspect of the configuration that a native speaker of the language says is appropriate. The semantic accuracy was 85.1%.

6 Discussion

This paper described a method for recovering the denotational meaning of a word, i.e. $\texttt{utter-W}(\Psi, \mathbf{x})$, given a set of sensory observations, each labeled according to whether it co-occurred with an utterance containing the word, i.e. $\texttt{hear-W}(\Psi, \mathbf{x})$. It was shown that $\texttt{hear-W}(\Psi, \mathbf{x})$ is a linear function of $\texttt{utter-W}(\Psi, \mathbf{x})$ where the parameters of the transform are determined by the level of shared attention and the background frequency of W. Given two weak assumptions about the form of $\texttt{utter-W}(\Psi, \mathbf{x})$, these parameters can be recovered and the transform inverted. The use of mutual information and randomization testing to identify the particular sensor group that captures a word's meaning was described. It is therefore possible to identify the denotational meaning of a word by simply observing the contexts in which it is and is not used, even in the face of imperfect shared attention and homonymy.

References

Paul R. Cohen. 1995. *Empirical Methods for Artificial Intelligence*. The MIT Press.

Eugene S. Edgington. 1995. *Randomization Tests*. Marcel Dekker.

Tim Oates. 2001. *Grounding Knowledge in Sensors: Unsupervised Learning for Language and Planning*. Ph.D. thesis, The University of Massachusetts, Amherst.

W. V. O. Quine. 1960. *Word and object*. MIT Press.

Terry Regier. 1996. *The Human Semantic Potential*. The MIT Press.

Eric J. Stollnitz, Tony D. DeRose, and David H. Salesin. 1996. *Wavelets for Computer Graphics: Theory and Applications*. Morgan Kaufmann.