Toward Natural Language Interfaces for Robotic Agents: Grounding Linguistic Meaning in Sensors

Tim Oates, Zachary Eyler-Walker and Paul R. Cohen
Computer Science Building
University of Massachusetts
Amherst, MA 01003-4610
{oates,zwalker,cohen}@cs.umass.edu

Even highly autonomous agents must be told what we want them to do. Ideally, we could communicate our goals to agents by talking with them, rather than encoding our goals in agent-specific knowledge representations. This paper explores one problem that lies in the way of this communicative ideal for robotic agents - the acquisition and grounding of linguistic meaning. How is it possible for a robot to identify and represent features of the physical world, as mediated by its ability to sense the world, that are relevant to understanding natural language utterances?

This paper describes the results of an experiment in which human subjects generated unrestricted natural language utterances to describe the activities of a Pioneer-1 mobile robot. A combination of word clustering applied to the utterances and a common subsequence algorithm applied to the time series of sensor values recorded by the robot made it possible for the Pioneer-1 to identify and represent a variety of features of its environment to which the utterances referred.

One problem we face is that natural languages encode features of the world at various levels of abstraction. For example, consider the difference between asking a robot to *push* an object and asking a robot to *shove* an object. The meanings of push and shove are similar in that both involve *contact* between two entities and the application of *force* by one entity to the other. However, the meanings of these words differ in the magnitude and duration of the force that is applied. Depending on the level of detail considered, these two words have either the same semantic features or similar, but different, features.

This paper investigates how robotic agents can identify linguistically relevant semantic features given natural language utterances and sensory access to the contexts in which they occur. We assume that the meanings of words are learned in an associationist manner as proposed by John Locke [2], through repeated exposure to utterances of a word in the presence of its referent, and describe an algorithm for performing such associationist learning with the sensory information available to a mobile robot.

Learning driven by the occurrence of individual words yields highly specific semantic features, i.e. the meanings if specific words. To identify more abstract semantic features we take advantage of the relationship between meaning and syntax – words with similar meanings tend to appear in the same syntactic constructions. For example, we would expect "book" and "newspaper" to appear in similar sentences more so than "book" and "carburetor". The agent is assumed to have no innate knowledge of syntax, and instead leverages the weak information about syntax available in word co-occurrences.

Given a corpus of sentences in a language, similarity of context (i.e. the surrounding words) can be used to hierarchically cluster words. Clusters correspond to sets of words that are similar both syntactically and semantically by virtue of the relationship between syntax and semantics. We use the clustering technique described in [1] which initially forms one cluster per unique word in the corpus, and proceeds to merge clusters bottom-up by choosing to merge the pair of clusters that results in the smallest loss of mutual information between clusters. The leaves of the hierarchy formed in this way are individual words, and movement up the hierarchy leads to clusters containing increasingly many words whose shared semantic features are necessarily more abstract. Locke essentially proposed using associationist mechanisms to identify semantic features at the leaves of this hierarchy, but they can also be used to identify the features shared by words further up as well.

Given a word cluster, $\mathcal{C}$, how might a robotic agent identify the semantic features associated with that cluster? For the sake of concreteness, assume the agent is a mobile robot and that its knowledge of the physical

environment is provided by a set of sensors. We assume that the presence of a word's referent in the physical environment induces a pattern, $\mathcal{P}$, in the time series produced by the robot's sensors. Let $p(\mathcal{P}|\mathcal{C})$ be the probability of the pattern occurring in the sensor data gathered when a member of $\mathcal{C}$ is uttered, and let $p(\mathcal{P}|\overline{\mathcal{C}})$ be the probability of the pattern when a member of $\mathcal{C}$ is not uttered. Under the reasonable assumption that words are uttered more frequently when their referent is present than when it is absent, it will be the case that $p(\mathcal{P}|\mathcal{C})$ is significantly different from $p(\mathcal{P}|\overline{\mathcal{C}})$.

Each time a word $w \in \mathcal{C}$ is uttered, the robot can record the values of its sensors over a window of time centered on the occurrence of the word. The result is a set of sensor time series that co-occurred with utterances of words in $\mathcal{C}$. The referent of w, and thus $\mathcal{P}$, may appear before, after or at the same time that w is uttered, and the relative timing of the two may change from one utterance to another. There is initially a total lack of knowledge concerning the location and the nature of $\mathcal{P}$ in the individual time series. The task facing the agent is to identify $\mathcal{P}$ given a set of time series gathered in this manner, and thereby to identify the semantic features associated with $\mathcal{C}$. We have developed an algorithm for solving this problem which is described in detail in [3], and which we use in the context of this work.

To evaluate the ability of the combined word clustering and subsequence discovery algorithm we ran an experiment in which a Pioneer-1 mobile robot was filmed engaging in 41 different activities, such as spinning in place or moving into a box or picking up small objects. Eight human subjects watched the film and wrote sentences describing the robot's behavior. They were instructed to begin all sentences with "The robot ..." and to keep the sentences as simple as reasonably possible, perhaps as though they were speaking to a young child or playing a text-adventure. Otherwise the sentences they produced were unrestricted.

Given the cluster hierarchy constructed from the eight human subjects' natural language sentences, the next task was to identify the semantic features of word clusters. Because words are typically uttered more frequently in the presence of their referent than in its absence, patterns in the robot's sensors that occur significantly more frequently in the presence of words in a cluster than in their absence are deemed to be semantic features of the cluster. That is, the problem of identifying linguistically relevant semantic features is cast in terms of identifying distinctive subsequences in the robot's sensor data.

During the filming of each scene the values produced by the robot's sensors were recorded at a rate of 10Hz. Given a word cluster, these time series were separated into two sets based on whether any of the human subjects used a member of the cluster when describing the associated scenes. These two sets of time series were then used to identify patterns in the robot's sensors (via the algorithm described in [3]) that occurred more frequently in the presence of words in the cluster than in their absence. This procedure was applied at every node in the word cluster hierarchy.

As an example of the type of results we obtained, the semantics of the word *pushed* were identified by looking for frequent subsequences in a subset of the robot's sensors containing the gripper state, the status of the break beam between the gripper paddles, and the status of the bump sensors on the tips of the gripper paddles. Three different frequent subsequences were identified in this case:

- gripper down, break beam on, bump switch off
- gripper down, break beam off, bump switch on
- gripper up, break beam off, bump switch on

These three configurations of the robot's sensors correspond to the following situations:

- pushing a small object between the robot's grippers on the floor
- pushing a large object that will not fit between the gripper paddles with the gripper down
- pushing an object with the gripper up and closed

There are three distinct ways that the robot can push objects, each of which was described with the word *pushed* by at least one subject, and the sensor time series that result from these situations were all identified as frequent subsequences.

References

[1] Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, 1992.

[2] John Locke. *An Essay Concerning Human Understanding.* Oxford : Clarendon Press, 1975. Original work published in 1690.

[3] Tim Oates. Identifying distinctive subsequences in multivariate time series by clustering. In *Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*, pages 322–326, 1999.