Toward Fast Mapping for Robot Adjective Learning

Allison Petrosino Kevin Gold

Wellesley College Dept. of Computer Science
Wellesley, MA 02482
allisonpetrosino@alum.wellesley.edu kevin.gold@gmail.com

Abstract

Fast mapping is a phenomenon by which children learn the meanings of novel adjectives after a very small number of exposures when the new word is contrasted with a known word. The present study was a preliminary test of whether machine learners could use such contrasts in unconstrained speech to learn adjective meanings and categories. Six decision tree-based learning methods were evaluated that use contrasting examples in order to work toward an adjective fast-mapping system for machine learners. Subjects tended to compare objects using adjectives of the same category, implying that such contrasts may be a useful source of data about adjective meaning, though none of the learning algorithms showed strong advantages over any other.

Introduction

The average American or British high school graduate knows, at a conservative estimate, 60,000 English words. The idea of learning and remembering so many words seems like a staggering feat—all the more so when you consider that this would mean that on average, a person must learn 10 new words each day up until that point (Bloom, 2002). Children who are 8-10 years old have been shown to learn at a higher than average rate (about 12 words each day) simply by going about their business as children (Anglin et al., 1993). It has further been shown that young children learn the meanings of new words much more quickly than adults do and with little explicit instruction (Newport, 1990).

This rapid word learning is called "fast mapping" in the child development literature (Carey and Bartlett, 1978; Bloom 2002) and is thought to crucially rely on contrast between examples. In the classic experiment on "fast mapping," children instructed to "point to the *chromium* tray, not the green tray" were often able to successfully deduce that "chromium" was a color word, and that it was the particular color implied by the statement (despite no pointing gestures), and could remember these facts weeks after the initial experiment (Carey and Bartlett, 1978).

The experiments presented here represent work that aims toward this kind of fast, situated learning for robots. Human subjects were asked to describe pairs of objects to a robot in a manner similar to the classic fast mapping questions. We then examined both a human-robot interaction (HRI) question as well as a machine learning question. On the human-robot interaction side: are these contrasts generally presented in a same-category manner ("drive to the red one, not the blue one"), or do the adjectives cross categories (blue/far)? On the machine learning side: can the contrasting examples provided by people help a decision-tree based learner to learn the meanings of words?

The machine learning problem of identifying definitions of words implicitly from just a few examples is rendered difficult by the fact that the robot may sense many kinds of properties, but only a few matter for a word's definition. The problem is not specific to machines, but is inherent in the problem (Wierzbicka, 1986). For example, a young learner told that a cow grazing in a field is "fangorious" does not know whether this word refers to the size of the cow, the number of legs it has, or the fact that it is spotted. But if people tend to contrast objects in a very formulaic manner, such as staying within category for contrasts, a machine learner could take advantage of any regularities to learn faster.

Deb Roy's CELL system (2002) and Chen Yu's gaze-tracking word learning system (Yu and Ballard, 2004)

both learned prototype definitions for nouns and color words based on object shape and color, but did not attempt to learn multiple adjective types, nor learn from explicit contrast. Terry Regier built a system that assigned spatial preposition labels to movies of a figure moving relative to a ground object, treating examples of prepositions as strong positive examples and all non-target words weak negative examples, but implicitly assumed that words generally referred to the same category of spatial language (1996). Stefanie Tellex later built a system that learned spatial routines using annotations of real video, with full sentences and phrases as input (Tellex and Roy, 2009). A great deal of modern word learning AI does not attempt to ground word meanings in sensory experience at all, but simply finds which words in text tend to co-occur (Landauer and Dumais, 1997). The present work is most similar to the TWIG system for learning word meanings from implicit contrast (Gold, 2009), but TWIG did not attempt to deal with the problem of determining which adjectives belong to the same category, instead assuming that all words belonging to a particular part of speech were mutually exclusive. Related work has attempted to determine whether adjective categories can be inferred automatically without using explicit contrast (Gold and Petrosino, 2010).

The next section describes and evaluates six decision tree-based methods that attempt to map meanings to color, size, and distance adjectives. Section 3 contains a discussion of the way contrast information was used in this work and speculates about its usefulness in more general terms. Section 4 provides some concluding remarks about the potential of these methods for use in future work as well as a reflection on the contribution of this work.

Methods

Real-world sensor data for this experiment was obtained using a Surveyor SRV-1 Blackfin robot (Figure 1). The SRV's camera and lasers were used to collect data about the color, the maximum height and width, and the distance scaled for height of objects in its environment.

Color information was measured in the YUV color space—a format that separates color information into luminance (Y) and chrominance (U and V) values. Not only has YUV been shown to lead to better segmentation of objects from their background than many other common color spaces, but luminance/chrominance color spaces also provide information that is more similar to the visual information taken in by the human retina than do other kinds of color spaces, RGB being one well-known example (Kumar, 2002; Livingstone, 2002).

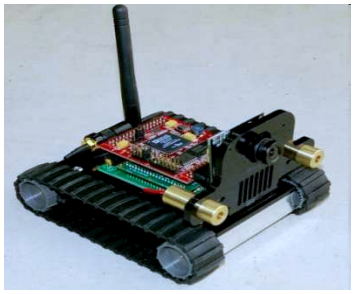

Figure 1: The Survyor SRV-1 Blackfin robot.

Thirty-two different objects were described in this experiment. Books, candles, and foam letters of varying heights, widths, and colors (red, green, orange, yellow, blue, and white, mostly monochromatic, though some objects had text or small blocks of other colors), and placed at three different distances (approximately 5, 25, and 45 cm) from the SRV were collected. Objects were chosen in order to avoid correlations between unrelated attributes (e.g., both tall and short green objects were chosen).

In order to obtain feature information for the objects, they were placed individually in front of the SRV's left laser and segmented from the background image using a depth-first search in which the laser was assumed to be pointing at the object and pixels whose color was closer to the color at the laser's position than a given threshold were added to the object. The YUV values assigned to the object also came from the color value at the laser's position. The distance from the SRV to the object was calculated using an empirically derived equation relating the position of the laser to object distance. Height and width values were simply the difference between the minimum and maximum x and y pixel values, scaled by distance.

Four Wellesley College students were recruited to perform three object description tasks in order to 1) determine which adjectives were likely to be used for instructing the robot, 2) obtain a consensus on descriptions to be used as ground truth for each object, and 3) ensure that the kinds of contrasts that people naturally make are useful for fast-mapping word learning. The data from three of the participants were used as a training set for building all

decision trees and the fourth participant's data was kept aside to be used as a test set.

In Task 1, designed to generate a list of size, distance, and color words for the SRV to learn and to determine which categories of adjectives were most commonly used to describe objects, participants sat at a table with the SRV placed directly in front of them. Objects were placed in front of the SRV in groups of 3 to 5. Participants were then asked to freely describe the objects without referring to any text on the objects (many of the objects were books). From this task, a list of words was generated for the robot to potentially learn, shown in Figure 2, divided into the three adjective categories for which the SRV takes in sensor information. Participants used an average of 2.8 adjectives to describe each object. At least one color word was used for each object by all of the participants; so in this task, color was the most common category used for free descriptions.

Adjective Category	Words Obtained from Free Descriptions
Color	red, yellow, blue, green, orange, white, salmon, maroon, brown
Size	small, big, tall, wide, little, short, thick
Distance	close, far, near

Figure 2: The list of all words used by participants.

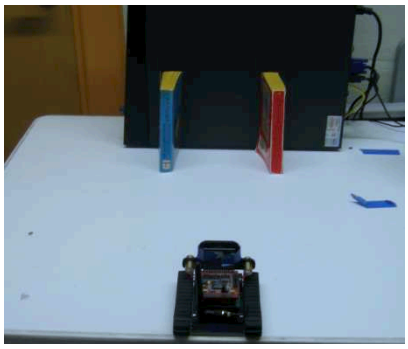

Figure 3: The experimental setup for task 3.

In Task 2, the words from the list created in the previous experiment were divided into pairs of opposites (color words were excluded) as follows: tall/short, big/little, large/small, wide/skinny, thick/thin, near/far, and close/far. In this experiment, participants were shown each of the 32 objects individually and asked, "What color would you say this object is?" The color word given for each object in this task matched one of the color words given for the object in Task 1. Then, for each pair of opposites, participants were asked, "Would you say that this object is ____, ____, or neither?" The descriptions obtained in this experiment served as a ground truth for determining whether or not the descriptions produced by the decision trees were correct. If "neither" was chosen, neither of the labels was used for the ground truth.

Task 3 was meant to determine whether people normally tend to contrast objects along a single dimension ("orange," not "blue"), which would make learning by fast mapping feasible outside of an experimental setting, or whether contrasts were commonly made across different adjective categories ("orange," not "short"). Participants were seated at a table with the SRV directly in front of them while pairs of objects differing along multiple dimensions were presented. Each object appeared in only one pair. The experimental setup for this task is shown below in Figure 3. Participants were then prompted with the following instruction: "How would you tell the robot to drive to one of the objects in front of it, giving instructions in the form, 'Drive to the ____ one, not the ____ one'? You may use only one adjective to describe each object."

The words used by subjects in task 3 and the 6 properties of the objects as measured by the robot (3 YUV features, distance, height, width) formed the training data for several different varieties of a greedy decision tree learning algorithm (Quinlan, 1986) that attempted to learn the definitions of each adjective.

One method attempted was a standard multiclass decision tree, in which only one word could be produced for each object. A second method added the difference in property values between the two contrasted objects as an extra six features, so that if **X** and **Y** were the feature vectors for the two objects, then **[X - Y X]** and **[Y - X Y]** were used as the new feature vectors for each example. A third method used the strategy of (Regier, 1996) and created a separate binary valued (Yes/No) tree for each adjective, treating examples with that adjective as strongly weighted ($w = 4$) positive examples and all other words as weak negative examples. This method was tried with and without chi-square pruning, in which decisions were omitted from the tree if they did not refer to a difference that was statistically significant. Two experiments were also run in which the yes or no trees were constrained to only contain decisions that all referred to the same category (color, size,

or distance); this approach was tried in both a greedy fashion, in which the first decision determined the category that would be used for the tree, and in an optimal fashion, in which all three categories were used for each tree, and only the category with best results were used. Finally, a graph was made of the subjects' contrasting comparisons, in which the vertices were the adjectives and an edge connected two vertices if subjects contrasted the words; multiclass trees were then run on the strongly connected components of this graph, in the hope that connected components would each be words that shared a category.

Precision was measured according the proportion of adjectives that the trees would produce for an example that agreed with the subjects' forced choices in Task 2. Recall was measured as the proportion of adjectives that could be used to describe the object that the learner produced, again using Task 2 for ground truth. F-measure was calculated as the harmonic mean of precision and recall.

Results

In total, 59 contrasts out of 64 total trials were within a single adjective category and only two of the four participants contrasted objects across categories. In 3 of the 5 cross-category contrasts, two different size dimensions were mixed, contrasting “tall” with “small” (strangely, both participants made this particular contrast) and “wide” with “short.” The remaining mixed-category contrasts came from one participant who contrasted “blue” with both “tall” and “near.” Again, color was the most common category used to describe words, with 36 out of the 64 contrasts within the color category.

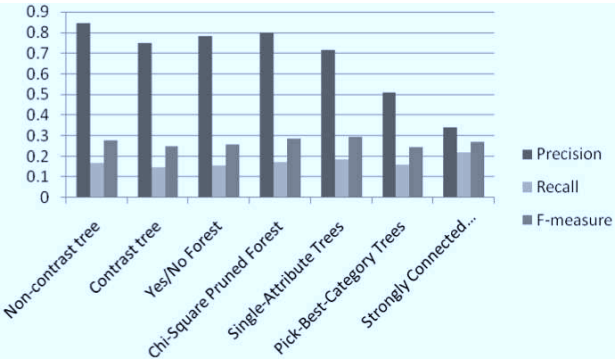

Figure 4: Machine learning results on the data for task 3.

The results of the decision tree learners are summarized in Figure 4. In general, it was much easier to achieve precision than recall in this setting because for all methods besides the strongly connected components method, the negative examples were all implicit, and occasionally wrong. For example, a red object labeled "wide" might count as a weak negative example for "red." This led the systems to be conservative in the words they produced. The strongly connected components method may benefit from a larger data set.

Conclusions

These preliminary experiments with human subjects show that there is merit to the assumption that people tend to contrast objects within categories instead of across categories, even when speaking to a robot. Thus, theoretically, machine learning algorithms should be able to learn words from people faster than if the examples had not been presented in contrast to each other, since this within-category constraint is an extra source of information. Why, then, did this not appear to be useful?

One possible explanation is that the experiment here was trying to accomplish two things at once -- both establish that people tend to use within-category contrasts, and also use this same data to perform machine learning. However, to establish the fact of within-category comparisons, object comparisons had to be used in which objects varied in more than one dimension -- a setting in which even children have a difficult time telling what a new adjective is referring to (O' Grady, 2000). However, it is useful to note that a robot using continuous-valued sensors will *always* encounter this problem, unless the sensed values are somehow quantized or some notion of salient difference is built in, because it is unlikely that any two continuous-valued features in the contrasted objects will be exactly the same. Our future approaches to this problem may attempt to take the magnitude of differences between features into account; the additional contrast features were a first pass at this, but were apparently not sufficient.

It may be the case that this work is a better model for early adjective learning than might be desired. Children have a great deal of difficulty in learning their first word in a given adjective category, particularly their first color word. The inability to correctly map colors to the appropriate

color words is so pronounced that Charles Darwin speculated that children begin their lives colorblind (Darwin, 1877). In one study of early color word learning, 2-year-olds were shown a series of different red objects and for each one they were asked, "What color is this?" When the children responded correctly, they were praised and when they failed to do so, they were gently corrected. It took an average of 85 trials before children reliably labeled the objects that they were being shown as "red" (Rice, 1980). It seems that for children first learning the words in an adjective category, a very large number of examples are required before the categories and the words that describe them are reliably linked, but once this point is reached, it is easy to learn new words within that category, as in the case of fast mapping. In these experiments, learning algorithms had only 96 training examples to learn 19 adjectives across three different categories. Attempting each of the learning algorithms described in this paper with a much larger data set could very well show that the number of examples required before they are successful is similar to the number of examples of a word that small children require for the initial learning of adjectives.

Although it was expected that most of the words used in this task would be simple descriptive adjectives ("tall", for example), participants' descriptions were often much more complex. Comparative and superlative adjectives ("taller" and "tallest"); negations ("not tall"); the addition of the suffix "-ish" to mean "somewhat" ("tallish"); adverbs ("pretty tall," "very tall"); and overall scene descriptions ("everything in this group is tall") were, taken together, used as often as simple descriptive adjectives were. It is beyond the scope of this paper to determine whether or how to incorporate these more informal or colloquial descriptions into a robot word learning system. However, if it is the case that people talking about the objects in their environment tend to use these more complex descriptions, it certainly seems that a robotic learner able to obtain useful information from such exchanges would be able to learn quickly and with little explicit instruction.

Although many of the learning algorithms presented in this paper did not clearly benefit from the addition of contrasting input data for learning the meanings of adjectives, it is in many ways a first step (or, at the least, a useful negative result) toward creating a system for the fast mapping of adjectives for robot learners. Much of the literature that addresses machine learning of adjectives is limited because it does not use real-world sensor data. All of the data in this thesis came from naïve participants freely describing objects and the camera and laser of a Surveyor SRV-1 Blackfin robot measuring their attributes. The noisiness of real-world data may make learning more difficult, but a machine learner's reliance on its own sensor data to obtain information about objects in the world has the potential to free the researcher from creating training data and presenting feedback to the learner. Some related works have used visually grounded data for learning the meanings of color words, but this is the first work in which an attempt was made to map meanings onto adjectives across categories, and to tackle size and distance words (Gold, 2009; Steels and Belpaeme, 2005). It is thus a first attempt at a grounded approach for fast mapping by offering as input two words at once that are in contrast to each other. Direct contrasts are an incredibly rich source of information and their usefulness to children learning new words should encourage those working toward fast, casual word learning for robots to take advantage of them.

Acknowledgments

This work was supported by the Norma Wilentz Hess fellowship of Wellesley College.

References

Jeremy Anglin. Vocabulary development: A morphological analysis. *Monographs of the Society for Research in Child Development* 58(10):1–166, 1993.

Paul Bloom. *How Children Learn the Meanings of Words.* The MIT Press, Cambridge, Massachussets, 2002.

Susan Carey and Elsa Bartlett. Acquiring a Single New Word. *Papers and Reports on Child Language Development*, 15:17-29, 1978.

Charles Darwin. A Biographical Sketch of an Infant. *Mind*, 2(7):285-294, 1877.

Kevin Gold. Using Sentence Context and Implicit Contrast to Learn Sensor-Grounded Meanings for Relational and Deictic Words: the TWIG System. PhD thesis, Dept. of Computer Science, Yale University, 2008.

Kevin Gold, Marek Doniec, Christopher Crick, and Brian Scassellati. Robotic Vocabulary Building Using Extension Inference and Implicit Contrast. *Artificial Intelligence*, 173:145-156, 2009.

Kevin Gold and Allison Petrosino. Using Information Gain to Build Meaningful Decision Forests for Multilabel Classification. In *Proceedings of ICDL 2010*, Ann Arbor, Michigan, 2010.

Pankaj Kumar, Kuntal Sengupta, and Adrian Lee. A Comparative Study of Different Color Spaces for Foreground and Shadow Detection for Traffic Monitoring System. In *Proceedings of the 5th IEEE Conference on Intelligent Transportation Systems*, Singapore, 2002.

Thomas K. Landauer and Susan T. Dumais. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review*, 104(2):211-240, 1997.

Margaret Livingstone. *Vision and Art: The Biology of Seeing*. Harry N. Abrams, Inc., New York, 2002.

Elissa L. Newport. Maturational Constraints on Language Learning. *Cognitive Science*, 14(1):11-28, 1990.

William O'Grady. *How Children Learn Languages*. Cambridge University Press, Cambridge, England, 2000.

J. Ross Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1):81-106, 1986.

Terry Regier. *The Human Semantic Potential: Spatial Language and Constrained Connectionism.* The MIT Press, Cambridge, Massachussets, 1996.

Mabel Rice. *Cognition to Language: Categories, Word meanings, and Training*. University Park Press, Baltimore, Maryland, 1980.

Deb K. Roy and Alex P. Pentland. Learning Words from Sights and Sounds: A Computational Model. *Cognitive Science.* 26:113-146, 2002.

Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, Upper Saddle River, New Jersey, 2nd edition, 2003.

Stefanie Tellex and Deb Roy. Grounding Spatial Prepositions for Video Search. In *Proceedings of the Eleventh International Conference on Multimodal Interfaces*. Cambridge, Massachusetts, 2009.

Anna Wierzbicka. What's in a Noun? (Or: How Do Nouns Differ in Meaning from Adjectives?). *Studies in Language*. 10(2):353-389, 1986.

Frank Yates. Contingency Table Involving Small Numbers and the Chi Square Test. *Journal of the Royal Statistical Society (Supplement)*, 1:217–235, 1934.

Chen Yu and Dana H. Ballard. A multimodal learning interface for grounding spoken language in sensory perceptions. *ACM Transactions on Applied Perceptions*, 1(1):57–80, July 2004.