Towards Visually-Grounded Spoken Language Acquisition

Deb Roy
Massachusetts Institute of Technology
Media Laboratory
20 Ames Street, Cambridge, MA 02142
dkroy@media.mit.edu

Abstract

A characteristic shared by most approaches to natural language understanding and generation is the use of symbolic representations of word and sentence meanings. Frames and semantic nets are examples of symbolic representations. Symbolic methods are inappropriate for applications which require natural language semantics to be linked to perception, as is the case in tasks such as scene description or human-robot interaction. This paper presents two implemented systems, one that learns to generate, and one that learns to understand visually-grounded spoken language. These implementations are part of our on-going effort to develop a comprehensive model of perceptually-grounded semantics.

1. Introduction

Conversational robots and other multimodal situated spoken dialogue systems raise serious design challenges regarding the mappings between natural language semantics and perception. Building on our previous investigations into cognitively-inspired computational models of language acquisition [8, 7], this paper describes two spoken language systems which ground semantics in visual perception. In both systems, the vision-to-language mappings are learned using statistical learning algorithms and "show-and-tell" training by a human teacher.

Grounding is used to refer, in part, to the process of connecting language to referents in the language user's environment. In contrast to methods which rely on symbolic representations of semantics, grounded representations bind words (and sequences of words) directly to non-symbolic perceptual features. Crucially, bottom-up sub-symbolic structures are available to influence symbolic processing [5]. The systems presented in this paper exemplify our effort to develop grounded language generation systems that couple language semantics with visual representations.

2 Learning to Generate Spoken Descriptions of Visual Scenes

We have implemented a trainable system called Describer which learns to generate descriptions of visual scenes by example (a more detailed description of this system can be found in [6]). A growing number of applications such as automatic sports commentators, talking maps, and web site description systems require the translation of perceptual information into natural language descriptions. Our eventual goal is to build a general purpose system which can be trained by example to perform these kinds of tasks.

Natural language semantics in Describer are anchored in features extracted from synthetic visual scenes. Input to the system consists of visual scenes paired with naturally spoken descriptions and their transcriptions. A set of statistical learning algorithms extract syntactic and semantic structures which link spoken utterances to visual scenes. These acquired structures are used by a generation algorithm to produce spoken descriptions of novel visual scenes. Concatenative synthesis is used to convert output of the generation subsystem into speech. In evaluations of semantic comprehension by human judges, the performance of automatically generated spoken descriptions is found to be comparable to human-generated descriptions.

The problem of generating referring expressions has been addressed in many previous computational systems (cf. [2, 1]). Most language generation systems may be contrasted with our work in two main ways. First, our emphasis is on learning all necessary linguistic structures from training data. Although some previous work has addressed trainable generation (for example, [3]), our goal is to automate training of both lexical choice and grammatical constructions. A second difference is that we take the notion of grounding semantics in sub-symbolic representations to be a critical aspect of linking natural language to visual scenes. All lexical and grammatical knowledge acquired by Describer is ultimately tied to visual representations.

Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02)
0-7695-1834-6/02 $17.00 © 2002 IEEE

2.1 The Verbal Description Task

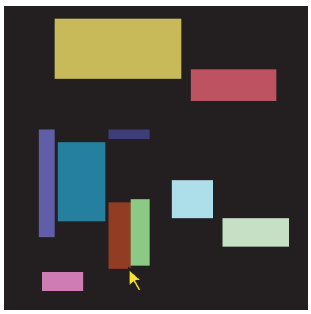

Figure 1. A typical scene processed by Describer. The arrow indicates the target object that must be verbally described.

The description task is based on images of the kind shown in Figure 1. The computer generated image contains a set of ten non-overlapping rectangles. The height, width, x-y position, and red-green-blue (RGB) color of each rectangle is continuously varying and chosen from a uniform random distribution. We addressed the following learning problem: Given a set of images, each with a *target object* and a natural language description of the target, learn to generate *syntactically correct, semantically accurate, and contextually appropriate* descriptions of objects embedded in novel multi-object scenes.

One basic problem is to establish the semantics of individual words. To bootstrap the acquisition of word associations, utterances are treated as "bags of words". Each word in an utterance may potentially be a label for any subset of co-occurring visual properties of the target. Thus the language learner must select relevant properties, that is, choose the subset of potential features which should be bound to a word. A second problem is to cluster words into word classes based on semantic and syntactic constraints. Word classes are a necessary first step in acquiring rules of word order. For example, before a language learner can learn the English rule that adjectives precede nouns, some primitive notion of adjective and noun word classes needs to be in place. A third problem is learning word order. We address the problems of learning adjective ordering ("the large blue square" vs. "the blue large square") and phrase ordering for generating relative spatial clauses. In the latter, the semantics of phrase order needs to be learned (i.e., the difference in meaning between "the ball next to the block" vs. "the block next to the ball").

Once word semantics and syntax have been learned, the system has at its disposal a grounded language model which enables it to map novel visual scenes to natural language descriptions. The language generation problem is treated as a search problem in a probabilistic framework in which syntactic, semantic, and contextual constraints are integrated.

2.2 Word and Grammar Learning

The 'perceptual system' of Describer consists of a set of feature extractors which operate on synthetic images. Each rectangle is described by a vector of 8 real-valued visual features: red, green, and blue color components, height-to-width ratio, area, x-position, y-position, and the ratio of the smaller dimension to the larger dimension. The training data consists of visual feature vectors of all objects in a scene paired with transcriptions of expressions referring to targets. Learning consists of six stages.

Stage 1: Word Class Formation

In order to generate syntactically correct phrases such as 'large red square' as opposed to 'red large square' or 'square red', word classes that integrate syntactic and semantic structure must be learned. Two methods of clustering words into syntactically equivalent classes were investigated. The first relies on distributional analysis of word co-occurrence patterns. The basic idea is that words which co-occur in a description are unlikely to belong to the same word class since they are probably labeling different properties of the target object. The second method uses shared visual grounding as a basis for word classification. A hybrid method which combines both methods led to optimal word clustering.

Stage 2: Feature Selection for Words and Word Classes

A subset of visual features is automatically selected and associated with each word. This is done by a search algorithm that finds the subset of visual features for which the distribution of feature values conditioned on the presence of the word is maximally divergent from the unconditioned feature distribution. Features are assumed to be normally distributed. The Kullback-Leibler divergence is used as a divergence metric between word-conditioned and unconditioned distributions. This method reliably selects appropriate features from the eight dimensional feature space. Word classes inherit the conjunction of all features assigned to all words in that class.

Stage 3: Grounding Adjective/Noun Semantics

For each word (token type), a multidimensional Gaussian model of feature distributions is computed using all observations which co-occur with that word. The Gaussian distribution for each word is only specified over the subset of features assigned to that word's class in Stage 2.

Stage 4: Learning Noun Phrase Word Order

A class-based bigram statistical language model is estimated (based on frequency) to model the syntax of noun phrases.

Stage 5: Grounding the Semantics of Spatial Terms

Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02)
0-7695-1834-6/02 $17.00 © 2002 IEEE

A probabilistic parser uses the noun phrase bigram language model from Stage 4 to identify noun phrases in the training corpus. Training utterances which are found to contain two noun phrases are used as input for this stage and Stage 6. Three spatial features between object pairs is introduced to ground spatial terms: the proximal angle, center-of-mass angle, and proximal distance [4]. The procedures in Stages 2 and 3 are re-used to ground spatial words in terms of these spatial features.

Stage 6: Learning Multi-Phrase Syntax

Multi-noun-phrase training utterances are used as a basis for estimating a phrase-based bigram language model. The class-based, noun phrase language models acquired in Stage 4 are embedded in nodes of the language model learned in this stage.

2.3 Pilot Acquisition Results

To train Describer, two human participants verbally described approximately 700 images. Figures 2-4 illustrate the results of the learning algorithm using this transcribed training corpus. The language model has a three-layer structure. Phrase order is modeled as a Markov model which specifies possible sequences of noun phrases and connector words, most of which are spatial terms (Figure 2). Two of the nodes in the phrase grammar designate noun phrases (labeled TARGET_OBJECT and LANDMARK_OBJECT). These nodes encapsulate copies of the phrase grammar shown in Figure 3. The semantics of relative noun phrase order have thus been learned and are encoded by the distinction of target and landmark phrases.

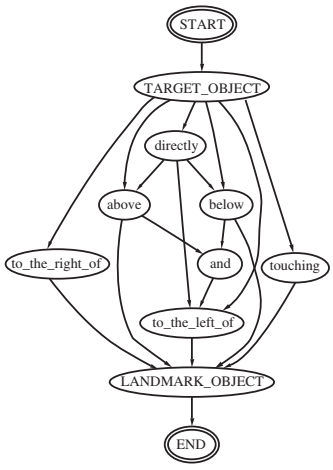

Figure 2. A grammar for combining object descriptions using relative spatial terms.

Each word class in Figure 3 is a result of learning Stage 1. Each word in the noun phrase language model is linked

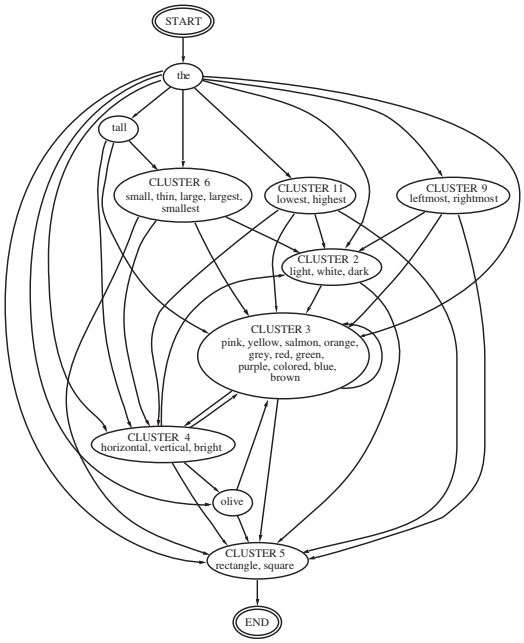

Figure 3. Noun phrase structure acquired by Describer.

to an associated visual model. The grounding models for one word class are shown as an example in Figure 4. The words 'dark', 'light' and 'white' were clustered into a word class in Stage 1. The blue and green color components were selected as most salient for this class in Stage 2. The ellipses in the figure depict iso-probability contours of the word-conditional Gaussian models in the blue-green feature space learned for each word in Stage 3. The model for 'dark' specifies low values of both blue and green components, whereas 'light' and 'white' specify high values. 'White' is mapped to a subset of 'light' for which the green color component is especially saturated.

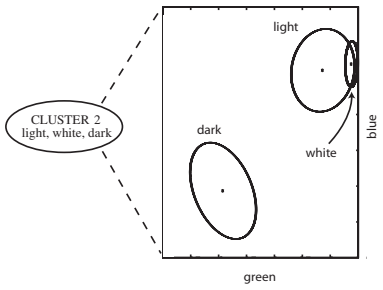

Figure 4. Visual grounding of words for a sample word class.

Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces (ICMI'02)
0-7695-1834-6/02 $17.00 © 2002 **IEEE**

2.4 Language Generation

A planning system uses the grounded grammar to generate semantically unambiguous, syntactically well formed, contextualized text descriptions of objects in novel scenes. A concatenative speech synthesis procedure is used to automatically convert the text string to speech using the input training corpus. The final output of the system are spoken descriptions of target objects in the voice of the human teacher. The planner works as follows:

Stage 1: Generate Noun Phrases

Using the noun phrase model as a stochastic generator, the most likely word sequence is generated to describe the target object, and each non-target object in the scene. Each word cluster specifies a probability distribution function for each word within the cluster. The Viterbi algorithm is used to find the most probable path through the graph (Figure 3) given a target object's visual features. The best path specifies a natural language referring expression.

Stage 2: Compute Ambiguity of Target Object Noun Phrase

An ambiguity score is computed based on how well the phrase generated in Stage 1 describes non-target objects in the scene. If the closest competing object is not well described by the phrase, then the planner terminates, otherwise it proceeds to Stage 3.

Stage 3: Generate Relative Spatial Clause

A landmark object is automatically selected which can be used to unambiguously identify the target. Stage 1 is used to generate a noun phrase for the landmark. The phrase-based language model is used to combine the target and landmark noun phrases.

Sample output is shown in Figure 5 for four novel scenes which were not part of the training corpus. Describer is able to generate descriptive utterances for each scene. Language generation is creative in the sense that the word sequences generated by Describer did not occur in the training set.

2.5 Evaluation

We evaluated spoken descriptions from the original human-generated training corpus and from the output of the generation system. Three human judges evaluated 200 human-generated and 200 machine-generated referring expressions. For each expression, judges were asked to select the best matching rectangle. Table 1 shows the evaluation results.

On average, the original human-generated descriptions were correctly understood 89.8% of the time. This result reflects the inherent difficultly of the task. An analysis of the errors reveals that a difference in intended versus inferred referents sometimes hinged on subtle differences in the speaker and listener's conception of a word. The average listener performance on the machine-generated descriptions was 81.3%, i.e., a difference of only 8.5% compared to the results with the human-generated set. An analysis of errors reveals that the same causes of errors found with the human set also were at play with the machine data.

In summary, Describer learns to generate verbal descriptions of objects in synthetic scenes with semantic accuracies comparable to human performance. The next section describes our efforts to address the converse problem: connecting verbal descriptions to objects in visual scenes.

The thin pink rectangle.

The highest vertical rectangle.

The dark green rectangle above the light green rectangle.

The dark purple rectangle touching the light purple rectangle.

Figure 5. Sample Describer output.

Table 1. Results of an evaluation of human and machine generated descriptions (chance performance is 10%).

Judge	Human-generated (% correct)	Machine-generated (% correct)
A	90.0	81.5
B	91.2	83.0
C	88.2	79.5
Average	89.8	81.3

3 Learning to Understand Descriptions of Objects in Video

Newt is a visually-grounded spoken language understanding system. The system processes spoken referring

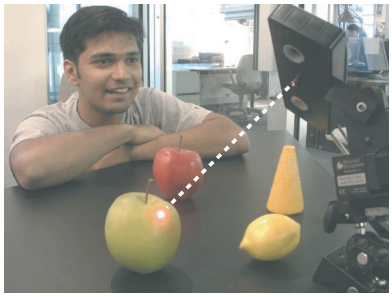

Figure 6. Newt is an interactive multimodal system which points to objects in response to verbal descriptions.

expressions such as "The green apple to the left of the cup" and locates the appropriate object in a visual scene. Newt is embodied in an active vision system mounted on a two degree-of-freedom pan-tilt base (Figure 6). A laser mounted in the device is used to point to objects on a table top in response to spoken utterances. This section provides an overview of Newt's implementation (more detailed descriptions are forthcoming).

3.1 Visual System

Newt's visual system tracks solid-colored objects placed on a table top in real-time. The system extracts object properties and inter-object spatial relationships which are passed to the language processing system. We model the color distribution of objects using mixtures of Gaussian distributions. Although all objects are constrained to be single-colored, shadow effects of three-dimensional objects necessitate the use of a mixture of Gaussian distributions. For each object used in the experiments, a color model is created by collecting training images of each object and manually specifying the region within each image that corresponds to the object. The Expectation Maximization (EM) algorithm is used to estimate both the mixture weights and the underlying Gaussian parameters for each object. K-means clustering is used to provide initial estimates of the parameters.

Shapes and colors are represented using multidimensional histograms of local color and geometrical features as introduced in [9]. These representations have been found to be robust to changes in lighting, scale, and in-plane rotation. An additional set of four shape related features is computed based on the bounding box of each object. These parameters are: height, width, height-to-width ratio, and area.

To enable the system to ground the semantics of spatial terms such as "above" and "to the left of", the set of spatial relations used in Described and described in [4] is measured between each pair of objects. The first feature is the angle (relative to the horizon) of the line connecting the centers of area of an object pair. The second feature is the shortest distance between the edges of the objects. The third spatial feature measures the angle of the line which connects the two most proximal points of the objects.

3.2 Speech Recognition

We have significantly extended our continuous speech recognition system [10] to support processing of interactive spoken language. The recognizer performs real-time medium vocabulary (up to 1000 word) recognition. We chose to develop our own recognizer in anticipation of non-standard decoder features which will be necessary to support rich integration with visual processing. The system uses a 24-band Mel-scaled cepstral front-end, continuous density HMM triphone acoustic sub-word models, and a back-off trigram statistical language model trained on a mixture of domain-specific and domain-independent data.

3.3 Language Learning by Show-and-Tell

Similar to Describer, Newt learns visually grounded language by 'show-and-tell'. During training sessions, Newt randomly points (using its laser pointer) to one of the objects in its view and waits for a human trainer to speak. This point-and-listen cycle is repeated with a variety of objects and object configurations as a rapid means of multimodal data collection. An on-line learning algorithm processes video-speech training pairs in order to acquire a visually-grounded grammar and lexicon which can be used for speech understanding.

Training examples consist of visual features of a target object and its spatial relation to other objects paired with transcriptions of spoken descriptions provided by a human trainer. As training examples arrive, statistical models of visual associations are spawned for each observed word. The association strength of a word to a particular visual feature is inspired by the methods developed for Describer. For each word, a record of all past visual contexts in which that word was observed are combined to estimate the parameters of Guassian distributions over all possible combinations of visual features. A background Gaussian model is computed using all past visual observations. The visual association strength of a word is computed as the Kullback-Leibler distance between the word-conditioned Gaussian and the background Gaussian model.

Word classes are formed using a hybrid strategy which combines distributional co-occurrence patterns of word usage with visual association similarities. For example, the words red and green are likely to be clustered because they rarely co-occur in the description of a single object, and

both terms have a strong visual association with color features (and no other visual features).

A hierarchical clustering process learns grammar fragments which encode the semantics of pairs of word classes or other grammar fragments. This hierarchical structure enables Newt to learn arbitrary depths of embedded phrase structure when provided with sufficient training data. In contrast, Describer is limited to a three level embedding of phrases, word classes, and words.

3.4 Situated Language Understanding

The language model acquired by Newt is suited to robust parsing of speech due to the distributed structure of the grammar fragments. To parse novel input, each grammar fragment attempts to explain the output of the speech recognizer. The set of fragments which best covers the input utterance is selected as the interpretation of the speech. Islands of words are sufficient for the understanding system to operate – a complete parse is not necessary nor expected. Each acquired word has an associated visual model (an expected Guassian distribution over a subset of visual features). The operation of the language-to-vision matching is best explained through an example. Consider the utterance "the ball above the red cup". The phrase red cup will be covered by a color-shape grammar fragment. A three-place grammar fragment will cover (the ball) (above) (the red cup). The visual model associated with ball will induce a probability mass distribution over all objects in the scenes, assigning most of the probability mass to round objects (assuming that ball has been learned correctly as labeling round shapes). Red cup will induce a second pmf over the objects which meet the conjunctive semantics of red and cup (conjunction is achieved at the moment by multiplying the pmf induced independently by each constituent term). Finally, 'above' picks out the object pair which individually have been assigned high probabilities, and whose spatial relation best fits the model associated with 'above'.

In a preliminary evaluation, we collected a dataset of 303 utterances from two trainers. Each utterance describes one object in a scene of four objects chosen from a collection of 10 objects in total, including objects with like shapes but different colors and vice versa. When trained on three of the sessions and evaluated on the fourth for all four sessions in turn, Newt achieves 82% accuracy in picking out the correct object, compared to a random baseline of 25%. Due to the small size of the dataset, we allowed Newt to use an example in the fourth session as a training example after Newt had selected an object for the example's utterance. Doing this increases performance by almost 10%, indicating that the dataset size is too small to achieve full performance. However, even this preliminary study shows that Newt does learn the correct visual groundings for words and their combinations.

4 Future Directions

We are currently expanding the set of visual features extracted from scenes to include topological features necessary for grounding spatial terms such as *through, in,* and *between*.

Not all natural language semantics seem to be grounded in perceptual representations. Some of the most frequent words in a young child's vocabulary include *I, you, my, yes, no, good*, and *want*. Motivated by such words (and their underlying concepts), we are investigating rudimentary models of intentionality and social reasoning to ground these non-perceptual concepts and make use of them in situated spoken dialogue.

5 Acknowledgements

Newt's grammar learning system has been developed by Peter Gorniak, and Newt's perceptual sub-systems have been built by Niloy Mukherjee.

References

[1] E. André and T. Rist. Generating coherent presentations employing textual and visual material. *Artificial Intelligence Review*, 9, 1995.

[2] R. Dale. *Generating Referring Expressions: Constructing Descriptions in a Domain of Objects and Processes.* MIT Press, 1992.

[3] P. Jordan and M. Walker. Learning attribute selections for non-pronominal expressions. In *Proceedings of ACL*, 2000.

[4] T. Regier. *The human semantic potential*. MIT Press, Cambridge, MA, 1996.

[5] D. Roy. Learning visually grounded words and syntax of natural spoken language. *Evolution of Communication*, 4(1), 2000/2001.

[6] D. Roy. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3), 2002.

[7] D. Roy. Grounded spoken language acquisition: Experiments in word learning. *IEEE Transactions on Multimedia*, In press.

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

[9] D. Roy, B. Schiele, and A. Pentland. Learning audio-visual associations from sensory input. In *Proceedings of the International Conference of Computer Vision Workshop on the Integration of Speech and Image Understanding*, Corfu, Greece, 1999.

[10] B. Yoder. Spontaneous speech recognition using hidden markov models. Master's thesis, Massachusetts Institute of Technology, Cambridge, MA, 2001.