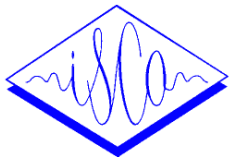

7th International Conference on Spoken Language Processing [ICSLP2002]
Denver, Colorado, USA
September 16-20, 2002

A SYSTEM THAT LEARNS TO DESCRIBE OBJECTS IN VISUAL SCENES

*Deb Roy**

The Media Laboratory
Massachusetts Institute of Technology
20 Ames Street, Cambridge, MA 02142

ABSTRACT

A spoken language generation system has been developed that learns to describe objects in computer-generated visual scenes. The system is trained by a 'show-and-tell' procedure in which visual scenes are paired with natural language descriptions. A set of learning algorithms acquire probabilistic structures which encode the visual semantics of phrase structure, word classes, and individual words. Using these structures, a planning algorithm integrates syntactic, semantic, and contextual constraints to generate natural and unambiguous descriptions of objects in novel scenes. The learning system is able to generalize from training data to generate expressions which never occurred during training. The output of the generation system is synthesized using word-based concatenative synthesis by drawing from the original training speech corpus. In evaluations of semantic comprehension by human judges, the performance of automatically generated spoken descriptions was comparable to human generated descriptions.

1. INTRODUCTION

A growing number of applications such as automatic sports commentators and talking maps require the translation of perceptual or sensory data into natural language descriptions. Most current approaches to this problem rely on manually created rules which encode domain specific knowledge. These rules are used for all aspects of the generation process including lexical and sentence frame selection. We present a trainable system called DESCRIBER which learns to generate descriptions of visual scenes by example (a more detailed description of this system is forthcoming [1]). This effort is motivated by our long term goal of developing spoken language processing systems which ground semantics in machine perception and action (for example, [2]).

We consider the problem of generating spoken descriptions of visual scenes to be a form of *language grounding*. Grounding refers, 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 must be available to influence symbolic processing [3].

Natural language semantics in DESCRIBER are visually grounded. 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. [4, 5]). 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. Jordan and Walker [6] also used machine learning to train a system to generate nominal descriptions of objects. This system learns to choose which logical combination of four attributes to use in describing objects. In comparison, the scope of what is learned by DESCRIBER includes attribute selection, syntactic structures and the visual semantics of words. 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. The Visual Translator system (VITRA) [7] grounds language generation in visual input (dynamic scenes from automobile traffic and soccer games). In contrast to our work, VITRA is not designed as a learning system. Thus porting it to a new domain would presumably be a arduous and labor intensive task. Our goal is to build a domain independent language learning system.

2. THE LEARNING TASK

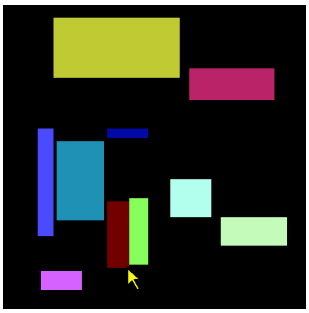

Fig. 1. A typical scene processed by DESCRIBER. The arrow indicates the target object that must be verbally described.

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

*This material is based upon work supported by the National Science Foundation under Grant No. 0083032.

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.

3. LANGUAGE ACQUISITION

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 clusters words which co-occur in similar visual contexts. This method uses shared visual grounding as a basis for word classification. We have found that a hybrid method which combines both methods leads to superior 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

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. Multi-noun-phrase utterances in our training corpus usually comprise a noun phrase describing the target object, followed by a spatial relation, followed by a *landmark* noun phrase. A typical utterance of this type is, 'The large square slightly to the left of the vertical pink rectangle'. Stable phrases are tokenized by iteratively finding word pairs with large bigram word pair probabilities. For example, 'to the left of' is converted to the token 'to_the_left_of'. Forward and reverse bigrams are used for this tokenization step (i.e., $P(w_t|w_{t+1})$ and $P(w_{t+1}|w_t)$). Tokenization enables visual semantics to be associated with whole phrases.

Any words in the training utterance which are not tagged as noun phrases by the parser are treated as candidate spatial terms. Three spatial primitives are introduced to represent inter-object relations. 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. 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.

3.1. Acquisition Results with a Pilot Training Corpus

To train DESCRIBER, two human participants verbally described approximately 700 images. Each spoken description was manually transcribed, resulting in a training corpus of images paired with utterance transcriptions. Figures 2-4 illustrate the results of the learning algorithm using this training corpus. The language model has a three-layer structure. At the highest level of abstraction (Figure 2), phrase order is modeled as a Markov model which specifies possible sequences of noun phrases and connector words, most of which are spatial terms. In addition to spatial terms, the system learned that 'touching' refers to configurations in which the proximal distance between object is 0. Transition probabilities have been omitted from the figure for clarity. 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. Note that at the phrase combination level (Figure 4), the semantics of relative noun phrase order are encoded by the distinction of target and landmark phrases. In other words, the system represents the fact that the first noun phrase describes the target and the second describes the landmark. This distinction is learned in Stage 6 (details of how this is learned are detailed in [1]).

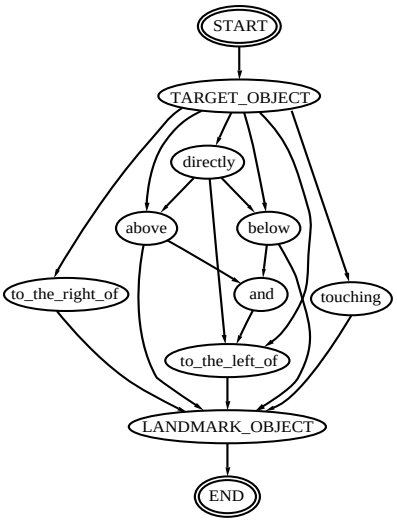

Fig. 2. A grammar for combining object descriptions using relative spatial terms.

Each word class in Figure 3 are a result of learning Stage 1. As in Figure 2, transition probabilities are not shown in Figure 3 due to space restrictions.

Each word in the noun phrase language model is linked 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.

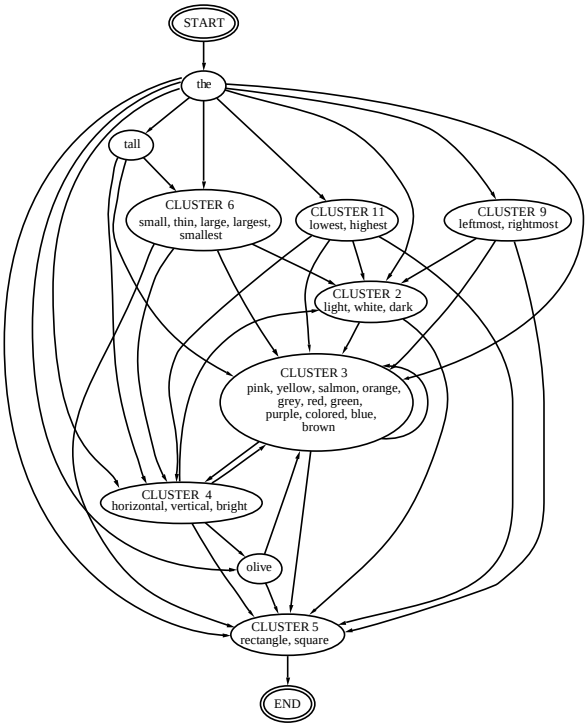

Fig. 3. Noun phrase structure acquired by DESCRIBER. The nodes labelled "TARGET_OBJECT" and "LANDMARK_OBJECT" in Figure 2 encapsulate copies of this structure.

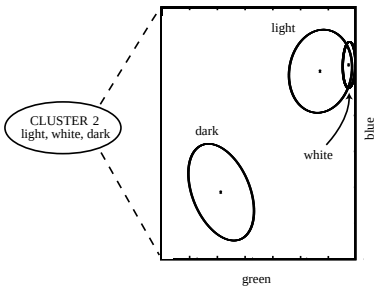

Fig. 4. Visual grounding of words for a sample word class. Each cluster in Figure 3 expands into a similar set of visual models defined interms of a set of visual features selected in learning Stage 2.

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 directly 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. In each scene, the target object is indicated with an arrow. Note that the descriptions take into account the relative context of each target object. In the lower two scenes, Stage 1 failed to produce an unambiguous noun phrase, so DESCRIBER generated a complex utterance containing a relative landmark. These descriptions represent DESCRIBER's attempt to strike a balance between syntactic, semantic, and contextual constraints.

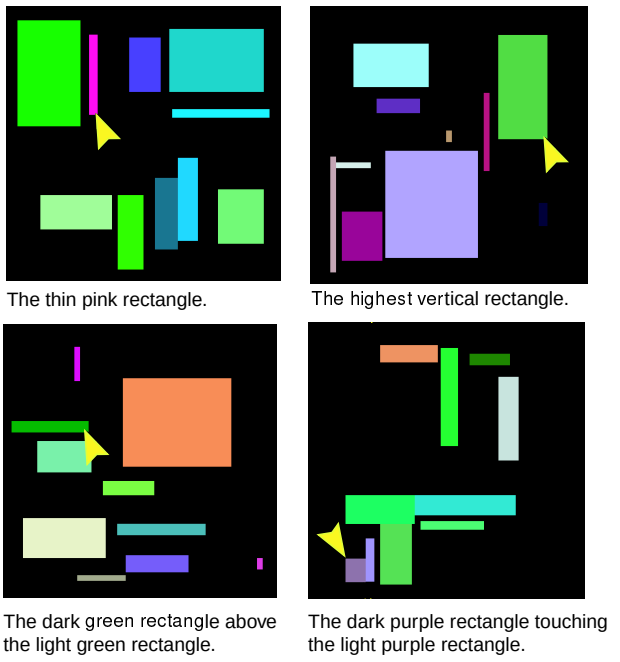

Fig. 5. Sample output generated by DESCRIBER for target objects indicated by arrows in the images. Relative spatial clauses are automatically generated to reduce ambiguity when needed (bottom two scenes).

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. For example the use of the terms "pink", "dark pink", "purple", "light purple", and "red" often lead to comprehension errors. In some cases it appears that the speaker did not consider a second object in the scene which matched the description he produced.

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

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 addition, we found that the system acquired an incorrect grounded model of the spatial term "to-the-left-of" which lead to several generation errors. This can easily be resolved by providing additional training examples which demonstrate proper use of the phrase.

6. CONCLUSIONS

We have presented an system which learns to describe objects in visual scenes using show-and-tell training. The learning method integrates distributional (syntactic) and semantic cues to create task-appropriate word classes. A hierarchical statistical language model is acquired in terms of these word classes which enables the language planner to generate natural language descriptions in response to novel visual input. Currently we are migrating the core structures and algorithms of DESCRIBER to a language *understanding* system which processes real-world visual input. This work furthers our long term efforts to develop systems which bridge the symbolic world of language processing to the non-symbolic world of visual representations.

7. REFERENCES

[1] Deb Roy, "Learning visually-grounded words and syntax for a scene description task," *Computer Speech and Language*, (in review).

[2] Deb Roy, "Grounded spoken language acquisition: Experiments in word learning," *IEEE Transactions on Multimedia*, 2001.

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

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

[5] Elisabeth André and Thomas Rist, "Generating coherent presentations employing textual and visual material," *Artificial Intelligence Review*, vol. 9, 1995.

[6] Pamela Jordan and Marilyn Walker, "Learning attribute selections for non-pronominal expressions," in *Proceedings of ACL*, 2000.

[7] Gerd Herzog and Peter Wazinski, "VIsual TRAnslator: Linking Perceptions and Natural Language Descriptions," *Artificial Intelligence Review*, vol. 8, pp. 175–187, 1994.