A TRAINABLE VISUALLY-GROUNDED SPOKEN LANGUAGE GENERATION SYSTEM

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. 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 output of the generation system is synthesized using word-based concatenative synthesis 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 require the translation of perceptual or sensory data into natural language descriptions. Most current approaches to this problem relies on manually created rules which encode domain specific knowledge. These rules are used for all aspects of the generation process including, for example, lexical selection and sentence frame selection. We present a trainable system called DESCRIBER (a more detailed description of this system can be found in [1]) which learns to generate descriptions of visual scenes by example. This work is motivated by our long term goal of developing spoken language processing systems which ground semantics in machine perception and action.

We consider the problem of generating spoken descriptions from visual scenes to be a form of *language grounding* [2, 3, 4]. Grounding refers 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 [2]. All symbolic representations are ultimately encoded in terms of representations of the machine's environment which are available to the machine directly through its perceptual system.

Semantics in DESCRIBER are visually-grounded. Input to the sytem 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 [5, 6]. Most previous 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 [7] also used machine learning to train a system which generates nominal descriptions of objects. The only aspect of this generation system which is trainable, however, is the choice of 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) [8] 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.

1.1. The Learning Problems

In this paper we consider learning problems in which each training example is comprised of (1) a natural language word sequence and (2) a vector of real-valued features which represents the semantics of the word sequence. We assume no prior knowledge about lexical semantics, word classes, nor syntactic structures.

A 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 features. Thus one problem facing the language learner is feature selection: choosing the subset of potential features which should be bound to a word. Once feature assignments have been made, statistical learning methods can be used to train classifiers which map words to ranges of values within those features. A second problem is to cluster words into word classes based on semantic and syntactic constraints. We assume that 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 presumably 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 the problems outlined above have been addressed, the system has at its disposal a grounded language model which enables it to map novel visual scenes into 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. SYSTEM DESCRIPTION

The description task is based on images of the sort 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. DESCRIBER addresses the following 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* referring expressions of objects embedded in novel multi-object scenes.

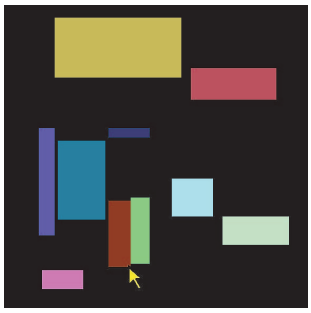

Fig. 1. A typical scene processed by DESCRIBER. The arrow indicates the target object.

The 'perceptual system' of DESCRIBER consists of a set of feature extractors which operate on synthetic images. In comparison to CELL, visual processing in DESCRIBER is trivially available since we have direct access to the source of the images (i.e., access to the program which generated the images). A set of visual attributes including shape, size, location, color, and brightness, is extracted from each rectangle in a scene. The features for the set of objects constitute the iconic representation of a scene. Learning in DESCRIBER 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 aspects of the scene. 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 an optimal clustering of words.

Stage 2: Feature Selection for Words and Word Classes

A subset of visual features is automatically selected and associated with each word. A search algorithm 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 has been found to reliably select word features in an 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 in Stage 2.

Stage 4: Learning Noun Phrase Word Order

A class-based bigram statistical language model is learned and models the syntax of noun phrases. The visually grounded word classes acquired in Stage 1 form the basis for this Markovian model of word order.

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. Utterances which are found to contain two noun phrases are used as input for this stage and Stage 6. Multi-noun-phrase utterances are usually of the form 'TARGET_NP [spatial relation] LANDMARK_NP', that is, 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.'. An automatic process based on bigram word pair probabilities is used to tokenize commonly occurring phrases (e.g., 'to the left of' is converted to the token 'to_the_left_of'). 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 in this stage to capture inter-object distance and angles. 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.

To train DESCRIBER, a human participant was asked to verbally describe approximately 500 images of the kind shown in Figure 1. Each spoken description was manually transcribed, resulting in a training corpus of images paired with utterance transcriptions[1]. Figures 3 and 4 illustrate the results of the learning algorithm on this training corpus. The language model has a three-layer structure. At the highest level of abstraction (left side of Figure 3), phrase order is modeled as a Markov model which specifies possible sequences of noun phrases and connector words, most of which are spatial terms. 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) and are diagrammatically linked by dashed lines

[1] A natural extension of this work is to integrate the acoustic word learning methods from CELL to replace this manual transcription step.

to the next level of the model. Note that at the phrase level, the semantics of relative noun phrase order are encoded by the distinction of target and landmark phrases. In other words, the system knows 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 can be found in [1].

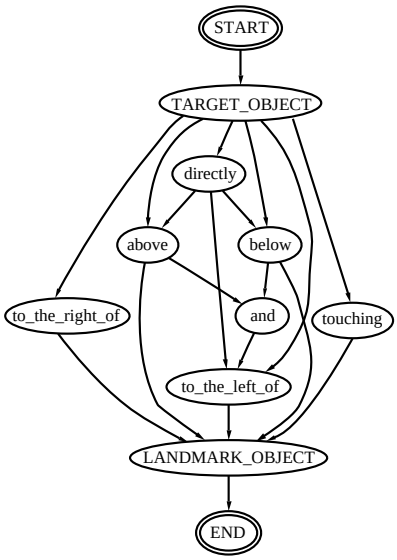

Fig. 2. Relative clause structure acquired by DESCRIBER.

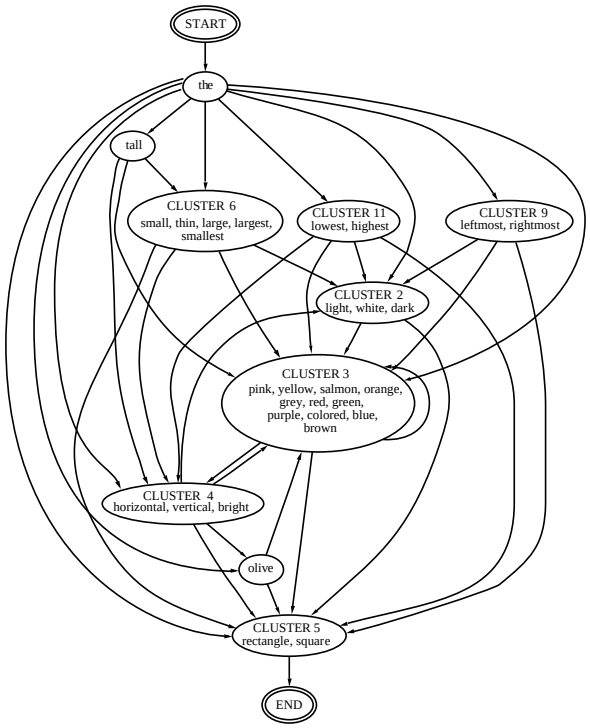

Fig. 3. Noun phrase structure acquired by DESCRIBER.

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 display isoprobability 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. In summary, the phrase level language model is grounded through two levels of indirection in terms of sensory features of the system.

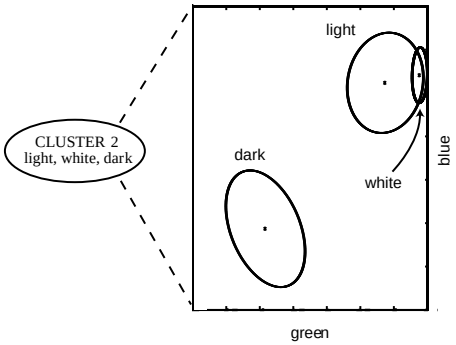

Fig. 4. Visual grounding of words for a sample word class.

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. In outline form, 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.

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. The Viterbi algorithm is used to compute the probability that each object in the scene matches the target phrase. If the closest competing object is not well described by the noun 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 from DESCRIBER 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

Table 1. Results of an evaluation of human and machine generated descriptions.

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

utterance containing a relative landmark. These descriptions represent DESCRIBER's attempt to strike a balance between syntactic, semantic, and contextual constraints.

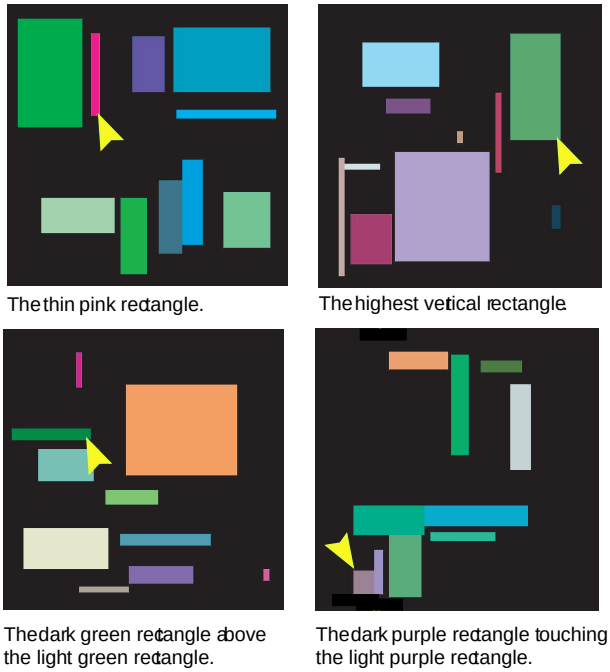

Fig. 5. Sample output generated by DESCRIBER for target objects indicated by arrows in the images. Relative spatial clauses are automatically generated to reduce ambiguity when needed (bottom two scenes).

3. EVALUATION

We evaluated spoken descriptions from the original human-generated training corpus and from the output of the generation system. Three human judges unfamiliar with the technical details of the generation system participated in the evaluation. Each judge evaluated 200 human-generated and 200 machine-generated spoken descriptions. All judges evaluated the same sets of utterances. Responses were evaluated by comparing the selected object for each image to the actual target object which was selected in order to produce the verbal description. Table 1 shows the results for both human-generated and machine generated results.

Averaged across the three listeners, the original human-generated descriptions were correctly understood 89.8% of the time. This result reflects the inherent difficultly of the rectangle 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 term. 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.

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. Differences in intended versus inferred meaning hinged on single descriptive terms. In some cases, an object was labeled using a descriptive term which was chosen mainly for its effect in reducing ambiguity rather than for its description accuracy. This lead at times to confusions for listeners. In addition, we also found that the system acquired an incorrect grounded model of the spatial term "to-the-left-of" which lead to some generation errors. This would easily be resolved by providing additional training examples which exemplify proper use of the phrase.

4. CONCLUSIONS

The results presented in this section demonstrate the effectiveness of the learning algorithms to acquire and apply grounded structures for the visual description task. The semantics of individual words, and the stochastic generation methods were able to produce natural spoken utterances which human listeners were able to understand with accuracies only 8.5% lower than original utterances spoken from the training corpus.

5. REFERENCES

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

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

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

[4] Deb Roy and Alex Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.

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

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

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

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