A Framework for Scene Generation from Spatial Descriptions

Dan Tappan; Advisor Roger Hartley

{dtappan,rth}@cs.nmsu.edu

New Mexico State University, Computer Science

Abstract

Text understanding by computers is generally limited to superficial processing of grammar and vocabulary only. Consequently, most systems cannot benefit from important contextual cues to narrow interpretations, reduce ambiguity, and so on, and their performance suffers accordingly. This framework focuses on bridging the gap between human and computer language processing by applying a knowledge-based solution to convert rudimentary text descriptions of static scenes into plausible visual interpretations.

Introduction

Natural language communicates descriptions of the world. Humans can decompose a complex visual scene into salient details, represent it with relatively few words, transmit it in written or verbal form, and then effortlessly reconstruct it with high fidelity. Very little information is actually stated, so humans rely heavily on commonsense knowledge and reasoning to fill in the gaps. Together, this explicit and implicit information helps the receiver build and manipulate a corresponding mental model of the scene, from which a picture can often be rendered.

This framework addresses three major problems in text understanding:

- *Underspecification*: obvious information is omitted
- *Vagueness*: interpretations depend heavily on context
- *Uncertainty*: multiple valid interpretations are possible

Input

A static scene depicts no movement and is described in terms of three components:

- The objects it contains.
- The spatial properties that the objects exhibit.
- The spatial relationships that hold between the objects.

The cactus is in front of the elephant and near it.
The big duck is near the elephant.

Figure 1: *Simple Scene Description*

Explicit Knowledge Representation

The three components can be represented conveniently as a directed graph:

Figure 2: *Semantic Network*

This contains only the information that was explicitly declared in the description and is thus incomplete and inadequate for any non-superficial interpretation.

Implicit Knowledge Representation

Filling the gaps in interpretation requires additional information not present in the description. This process corresponds to using our knowledge of the world that we acquire and refine throughout life.

Any computational solution requires a knowledge base of specialized facts and definitions. Here a greatly simplified form is organized hierarchically by generic concepts with spatial details:

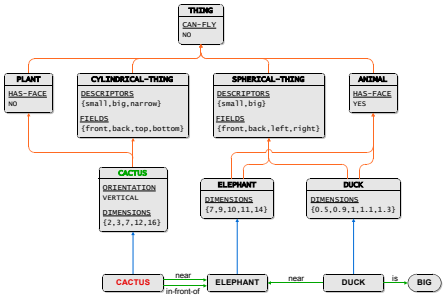

Figure 3: *Linked Knowledge Base*

Thus the generic concept `cactus` (in green) is defined by its own information, as well as that of its ancestor concepts `cylindrical-thing`, `plant`, and `thing`.

By linking the specific cactus (in red) in the semantic network to its generic concept, significant additional knowledge is available to augment the description.

Dimensional Reasoning

Reasoning over size depends on the objects; e.g., a duck is big in volume, but a forest is in area; or a huge duck has less size than a tiny elephant, etc. The knowledge base defines such contexts, as well as probabilistic value ranges for each dimension. For example, an elephant short (in height) but long (in depth) with average width may be 9±0.8 feet high, 11±0.7 long, and 5.5±0.7 wide:

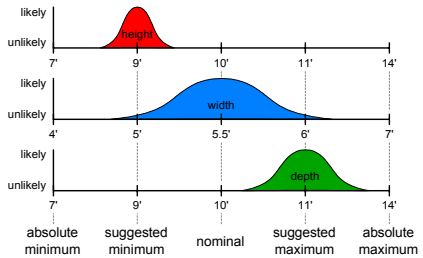

Figure 4: *Plausible Dimensions for an Elephant*

Positional Reasoning

Reasoning over position and direction also depends on the objects; e.g., the elephant can face the cactus, but not vice versa. Geometric regions associated with each object impose such constraints. In Figure 5a, the blue wedge defines where anything can appear in front of the elephant, and the red ring defines near it. The definition of *in front of* and *near* thus corresponds to where the regions overlap.

The geometry defines where other objects can appear, and the overlaid topology refines within that region which positions are more likely. The meshes in Figure 5b show the probability distributions.

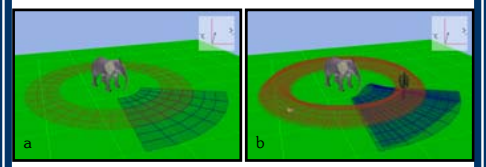

Figure 5: *Geometry and Topology Regions*

Output

The final interpretation appears in an interactive virtual world that can be viewed in various ways and used as a test-and-evaluation environment:

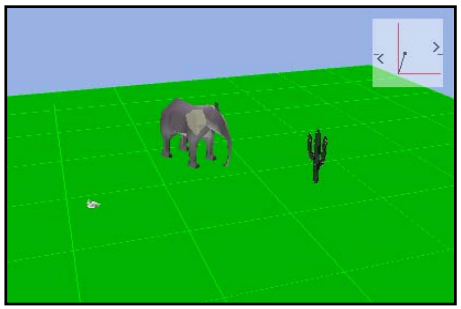

Figure 6: *Rendered Interpretation*

In addition, new details from this solution are inferred and inserted (in red) back into the original semantic network to augment its understanding even further:

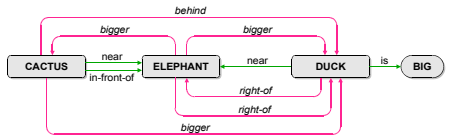

Figure 7: *Augmented Semantic Network*

Conclusion

This framework is part of a larger system for testing and evaluating stochastic simulations of these reasoning processes. The interactive visual environment allows individual issues to be isolated and tested formally.

As a work in progress, it has yet to produce definitive results. Nevertheless, preliminary data suggest that it successfully addresses a number of vexing problems in understanding spatial descriptions: underspecification and vagueness are handled by the knowledge base, and uncertainty is handled by the probabilistic nature of reasoning over the information it provides.

MC-04

This poster was printed by NM-BRIN, NCRR NIH Grant Number RR-16480