Knowledge-Based Spatial Reasoning for Scene Generation from Text Descriptions

Dan Tappan

College of Engineering, Idaho State University
921 S. 8th Ave., Stop 8060
Pocatello, ID 83209-8060
tappdan@isu.edu

Abstract

This system translates basic English descriptions of a wide range of objects in a simplistic zoo environment into plausible, three-dimensional, interactive visualizations of their positions, orientations, and dimensions. It combines a semantic network and contextually sensitive knowledge base as representations for explicit and implicit spatial knowledge, respectively. Its linguistic aspects address underspecification, vagueness, uncertainty, and context with respect to intrinsic, extrinsic, and deictic frames of spatial reference. The underlying, commonsense reasoning formalism is probability-based geometric fields that are solved through constraint satisfaction. The architecture serves as an extensible test-and-evaluation framework for a multitude of linguistic and artificial-intelligence investigations.

Introduction and Background

A simple description like *a large dog is in front of a cat and near a small tree* explicitly specifies only a tiny fraction of the details that a corresponding image contains. Most of the content comes from an implicit, commonsense, contextual understanding of the words. Such spatial reasoning, like most intelligent processes, is a difficult computational task to emulate despite its apparent, intuitive simplicity for humans (Herskovits 1986, Tversky 2000). What makes the problem especially troublesome is that computers lack our intangible knowledge of the world and powerful abilities to reason intelligently over it. This work addresses the primary aspects of these issues in terms of what to represent and how to represent it. It uses a simple representation of a description in conjunction with a knowledge base of relevant spatial details to define the declarative form of a valid solution. A constraint satisfaction algorithm then generates any number of corresponding interpretations with plausible positions, orientations, and dimensions for the objects.

Four knowledge-based spatial issues are the focus: *underspecification*, or the lack of complete details in a description, requires background knowledge to supply implicit information; *vagueness*, or the imprecise nature of descriptions, requires knowledge that defines a range of plausible interpretations; *uncertainty*, or the lack of commitment to a particular interpretation, requires knowledge of preferences over this range; and *context*, or the different interpretation of objects in certain combinations, requires knowledge to identify and interpret such patterns.

These issues are considered for three frames of spatial reference (Olivier and Tsujii 1994). The *intrinsic* (object-centered) frame generally applies to objects that have an accepted front, like *dog*. The *extrinsic* (environment-centered) frame and the *deictic* (viewer-centered) frames are generally the opposite case for objects without such a front, like *tree*. They correspond to the viewer's position being explicitly stated or loosely implied, respectively.

Knowledge Representation

A description consists of nouns, adjectives, prepositions, and various support words. The nouns refer primarily to animals and plants within a zoo scenario because they exhibit a variety of interesting and visually appealing spatial characteristics. The adjectives play a role in the contextually appropriate determination of size. The prepositions are 58 spatial relations for position (e.g., *in front*, *left*, *north*, *between*), distance (e.g., *inside*, *near*, *far*), and orientation (e.g., *facing toward*, *away from*, *north*).

Explicit Representation

The explicit knowledge in a description is represented with a semantic network of object nodes, attribute nodes, and directed relation arcs, which map closely to nouns, adjectives, and prepositions, respectively. For example, Figure 1a depicts the semantic network for *Loki is a small retriever; the tree is north of Loki; Loki is facing the tree.*

Implicit Representation

To understand the meaning of the description even superficially requires deeper analysis into what the objects are and how their spatial rules apply to them. This implicit, commonsense background knowledge is represented in a knowledge base that is similar to an inheritance hierarchy in object-oriented programming. It currently contains 108 concepts that either inherit their contents from their ancestors or define/override them. A simplified example appears in Figure 1b. Linking the semantic network to the knowledge base provides the objects with the appropriate rules for interpreting their position and orientation relations and dimension attributes.

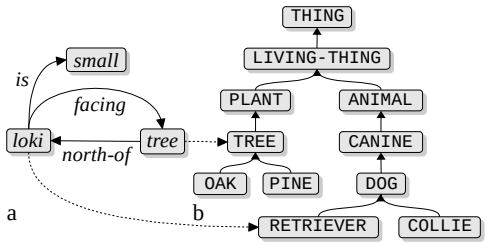

Figure 1: Semantic Network and Knowledge Base

Spatial Relations

Each spatial relation is associated with one or more circular, two-dimensional fields of 100 rings and 32 sectors that have two complementary parts (Tappan 2004, Yamada 1993, Gapp 1994, Olivier and Tsujii 1994, and Freska 1992). The *geometry* specifies where another object can and cannot appear with respect to the object in the center. Most relations use variants of the wedge and ring fields in Figures 2a-b. The *topography* overlays a probability distribution onto the geometry to specify preferences in placement, as Figures 2c-d show. Fields may also be combined with the standard logical operators *and*, *or*, *xor*, and *not* to represent compositional linguistic expressions like *in front of and far from*.

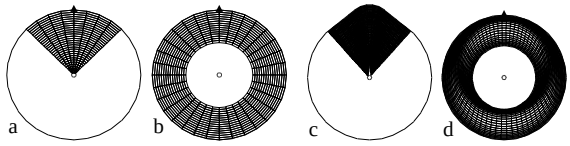

Figure 2: Geometry and Topography of Wedge and Ring

Spatial Reasoning

The intelligent, commonsense aspects of the spatial reasoning are actually performed above by establishing their contextually appropriate, qualitative constraints. Generating a solution from them is now a straightforward, mechanical process of quantitative constraint satisfaction using a greedy, backtracking strategy to generate and test positions and orientations for every pair of objects in a relationship.

Interactive Visualization

The graphical output is a three-dimensional, interactive world, in which the viewer can move to any vantage point and perspective. It is also possible to query the objects on their underlying representations and constraints, etc. Various display modes depict supporting details like the geometry and topography of the fields, as well as alternative solutions. Figure 3 renders *the dog is south of the tree and near the panther; the panther is to the right of the dog; and the elk is near the maple tree and near and facing away from the pond*.

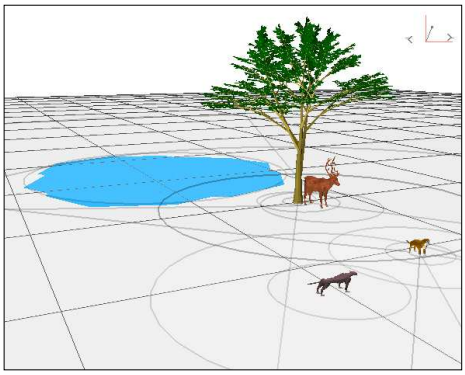

Figure 3: Sample Visualization

References

Freska, C. 1992. Using Orientation Information for Qualitative Spatial Reasoning. In Frank, A.; Campari, I.; and Formentini, U., eds. *Theories and Methods of Spatio-Temporal Reasoning in Geographic Space*, LNCS 639, Springer-Verlag, Berlin.

Gapp, K. 1994. Basic Meanings of Spatial Relations: Computation and Evaluation in 3D Space. In *Proceedings of AAAI-94*, 1393-1398. Seattle, WA.

Herskovits, A. 1986. *Language and Spatial Cognition: An interdisciplinary Study of the Prepositions in English*. Cambridge: Cambridge University Press.

Olivier, P.; and Tsujii, J. 1994. A computational view of the cognitive semantics of spatial prepositions. In *Proceedings of 32nd Annual Meeting of the Association for Computational Linguistics (ACL-94)*, Las Cruces, New Mexico.

Tappan, D. 2004. Knowledge-Based Spatial Reasoning for Automated Scene Generation from Text Descriptions. Ph.D. diss., New Mexico State University.

Tversky, B. 2000. Levels and structure of spatial knowledge. In *Cognitive Mapping: Past, present and future*, Kitchin, R.; and Freundshuh, S., eds. London and New York: Routledge.

Yamada, A. 1993. Studies on Spatial Description Understanding Based on Geometric Constraints Satisfaction. Ph.D. diss., University of Kyoto.