Spatial Relations in Text-to-Scene Conversion

Bob Coyne[1], Richard Sproat[2], and Julia Hirschberg[1]

[1] Columbia University, New York NY, USA,
{coyne, julia}@cs.columbia.edu,
[2] Oregon Health & Science University, Beaverton, Oregon, USA
rws@xoba.org

Abstract. Spatial relations play an important role in our understanding of language. In particular, they are a crucial component in descriptions of scenes in the world. WordsEye (www.wordseye.com) is a system for automatically converting natural language text into 3D scenes representing the meaning of that text. Natural language offers an interface to scene generation that is intuitive and immediately approachable by anyone, without any special skill or training. WordsEye has been used by several thousand users on the web to create approximately 15,000 fully rendered scenes. We describe how the system incorporates geometric and semantic knowledge about objects and their parts and the spatial relations that hold among these in order to depict spatial relations in 3D scenes.

1 Introduction

Spatial relations are expressed either directly or implicitly in a wide range of natural language descriptions. To represent these descriptions in a 3D scene, one needs both linguistic and real-world knowledge, in particular knowledge about: the spatial and functional properties of objects; prepositions and the spatial relations they convey, which is often ambiguous; verbs and how they resolve to poses and other spatial relations. For example, to interpret *apple in the bowl* we use our knowledge of bowls – that they have interiors that can contain objects. With different objects (e.g., *boat in water*), a different spatial relation is conveyed.

WordsEye [6] is a system for automatically converting natural language text into 3D scenes representing the meaning of that text. A version of WordsEye has been tested online (www.wordseye.com) with several thousand real-world users. We have also performed preliminary testing of the system in schools, as a way to help students exercise their language skills. Students found the software fun to use, an important element in motivating learning. As one teacher reported, "One kid who never likes anything we do had a great time yesterday...was laughing out loud."

WordsEye currently focuses on directly expressed spatial relations and other graphically realizable properties. As a result, users must describe scenes in somewhat stilted language. See Figure 1. Our current research focuses on improving the system's ability to infer these relations automatically. However, in this paper, we describe the basic techniques used by WordsEye to interpret and depict directly expressed spatial relations.

In Section 2 we describe previous systems that convert natural language text to 3D scenes and prior linguistic work on spatial relations. In Section 3 we provide an overview of WordsEye. In Section 4 we discuss the spatial, semantic and functional knowledge about objects used to depict spatial relations in our system. We conclude and describe other ongoing and future work in Section 5.

2 Prior Work

Natural language input has been investigated in some early 3D graphics systems [1][13] including the Put system [4], which was limited to spatial arrangements of existing objects in a pre-constructed environment. In this system, input was restricted to an artificial subset of English consisting of expressions of the form $Put(X, P, Y)$, where X and Y are objects and P is a rigidly defined spatial preposition. Work at the University of Pennsylvania's Center of Human Modeling and Simulation [2], used language to control animated characters in a closed virtual environment. CarSim [7] is a domain-specific system that creates animations from natural language descriptions of accident reports. CONFUCIUS [12] is a multi-modal text-to-animation system that generates animations of virtual humans from single sentences containing an action verb. In these systems the referenced objects, attributes, and actions are typically relatively small in number or targeted to specific pre-existing domains.

Spatial relations have been studied in linguistics for many years. One reasonably thorough study for English is Herskovits [9], who catalogs fine-grained distinctions in the interpretations of various prepositions.[3] For example, she distinguishes among the various uses of *on* to mean "on the top of a horizontal surface" (*the cup is on the table*), or "affixed to a vertical surface" (*the picture is on the wall*). Herskovits notes that the interpretation of spatial expressions may involve considerable inference. For example, the sentence *the gas station is at the freeway* clearly implies more than just that the gas station is located next to the freeway; the gas station must be located on a road that passes over or under the freeway, the implication being that, if one proceeds from a given point along that road, one will reach the freeway, and also find the gas station.

[3] It is important to realize that how spatial relations are expressed, and *what kinds of relations may be expressed* varies substantially across languages. Levinson and colleagues [11] have catalogued profound differences in the ways different languages encode relations between objects in the world. In particular, the Australian language Guugu Yimithirr and the Mayan language Tzeltal use absolute frames of reference to refer to the relative positions of objects. In Guugu Yimithirr, one can locate a chair relative to a table only in terms of cardinal points saying, for example, that the chair is north of the table. In English such expressions are reserved for geographical contexts — *Seattle is north of Portland* — and are never used for relations at what Levinson terms the "domestic scale". In Guugu Yimithirr one has no choice, and there are no direct translations for English expressions such as *the chair is in front of the table*.

Eye of the Beholder by Bob Coyne

Input text: The silver penny is on the moss ground. The penny is 7 feet tall. A clown is 2 feet in front of the penny. The clown is facing the penny.

No Dying Allowed by Richard Sproat

Input text: Eight big white washing machines are in front of the big cream wall. The wall is 100 feet long. The "No Dying Allowed" whiteboard is on the wall. The whiteboard is one foot high and five feet long. The ground is tile. Death is in front of the washing machines. It is facing southeast. Death is eight feet tall.

Fig. 1: Some Examples from WordsEye's Online Gallery

3 System Overview

Our current system is an updated version of the original WordsEye system [6], which was the first system to use a large library of 3D objects to depict scenes in a free-form manner using natural language. The current system contains 2,200 3D objects and 10,000 images and a lexicon of approximately 15,000 nouns. It supports language-based control of objects, spatial relations, and surface properties (e.g., textures and colors); and it handles simple coreference resolution, allowing for a variety of ways of referring to objects. The original WordsEye system handled 200 verbs in an *ad hoc* manner with no systematic semantic modeling of verb alternations and argument combinations. In the current system, we are instead adding frame semantics to support verbs more robustly. To do this, we are utilizing our own lexical knowledge-base, called the SBLR (Scenario-Based Lexical Resource) [5]. The SBLR consists of an ontology and lexical semantic information extracted from WordNet [8] and FrameNet [3] which we are augmenting to include the finer-grained relations and properties on entities needed for depicting scenes as well as capturing the different senses of prepositions related to those properties and relations.

The system works by first parsing each input sentence into a dependency structure. These dependency structures are then processed to resolve anaphora and other coreferences. The lexical items and dependency links are then converted to semantic nodes and roles drawing on lexical valence patterns and other information in the SBLR. The resulting semantic relations are then converted to a final set of graphical constraints representing the position, orientation, size, color, texture, and poses of objects in the scene. Finally, the scene is composed from these constraints and rendered in OpenGL (`http://www.opengl.org`) and

optionally ray-traced in Radiance [10]. The user can then provide a title and caption and save the scene in our online gallery where others can comment and create their own pictures in response. See Figure 1.

4 Spatial Relations

WordsEye uses SPATIAL TAGS and other spatial and functional properties on objects to resolve the meaning of spatial relations. We focus here on the interpretation of NPs containing spatial prepositions of the form "X-preposition-Y", where we will refer to X as the FIGURE and Y as the GROUND. For example, in *snow is on the roof*, *snow* is the FIGURE and *roof* is GROUND. The interpretation of the spatial relation often depends upon the types of the arguments to the preposition. There can be more than one interpretation of a spatial relation for a given preposition The geometric and semantic information associated with those objects will, however, help narrow down the possibilities.

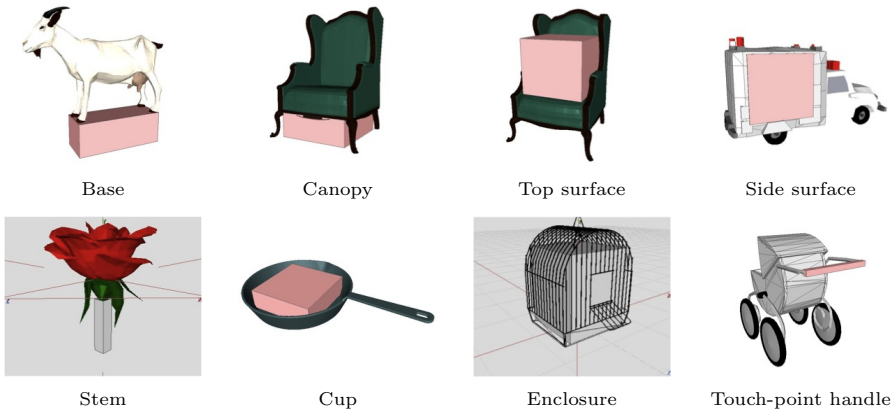

Fig. 2: Spatial Tags, represented here by the boxes associated with each object, designate regions of those objects used in resolving spatial relations. For example, the TOP SURFACE region marked on the seat of the chair is used in sentences like *The pink mouse is on the small chair* to position the FIGURE (*mouse*) on the GROUND (*chair*). See Figure 3 for the depiction of this sentence and several others that illustrate the effect of spatial tags and other object features.

The 3D objects in our system are augmented with the following features:

- IS-A: The lexical category to which the given object belongs.
- Spatial tags identifying the following regions: (See Figure 2)
  - CANOPY: A canopy-like area "under" an object (e.g., *under a tree*).
  - CUP: A hollow area, open above, that forms the interior of an object.
  - ENCLOSURE: An interior region, bounded on all sides (holes allowed).

    - TOP/SIDE/BOTTOM/FRONT/BACK: For both inner and outer surfaces.
    - NAMED-PART: For example, the hood on car.
    - STEM: A long thin, vertical base.
    - OPENING: An opening to an object's interior (e.g., doorway to a room).
    - HOLE-THROUGH: A hole through an object. For example, a ring or donut.
    - TOUCH-POINT: Handles and other functional parts on the object. For example, in *John opened the door*, the doorknob would be marked as a handle, allowing the hand to grasp at that location.
    - BASE: The region of an object where it supports itself.
- OVERALL SHAPE: A dominant overall shape used in resolving various spatial relations. For example, SHEET, BLOCK, RIBBON, CUP, TUBE, DISK, ROD.
- FORWARD/UPRIGHT DIRECTION: The object's default orientation.
- SIZE: The default real-world size of the object. This is also used in spatial relations where the FIGURE and GROUND size must be compatible. For example, *ring on a stick* versus **life-preserver on a pencil*.
- LENGTH AXIS: The axis for lengthening an object.
- SEGMENTED/STRETCHABLE: Some objects don't change size in all dimensions proportionally. For example, a fence can be extended indefinitely in length without a corresponding change in height.
- EMBEDDABLE: Some objects, in their normal function, are embedded in others. For example, fireplaces are embedded in walls, and boats in water.
- WALL-ITEM and CEILING-ITEM: Some objects are commonly attached to walls or ceilings or other non-upward surfaces. Some (e.g., pictures) do this by virtue of their OVERALL SHAPE, while for others (e.g., sconces) the orientation of the object's BASE is used to properly position the object.
- FLEXIBLE: Flexible objects such as cloth and paper allow an object to hang or wrap. For example, *towel over a chair*.
- SURFACE ELEMENT: Any object that can be part of a flat surface or layer. For example, a crack, smudge, decal, or texture.
- Semantic properties such as PATH, SEAT, AIRBORNE for object function.

Some of these features were used in earlier versions of our system [6]. Features we have added to the current version include: SURFACE ELEMENT, EMBEDDABLE, OVERALL SHAPE, LENGTH AXIS, SEGMENTED/STRETCHABLE. Other features, including (FLEXIBLE, OPENING, HOLE-THROUGH and various semantic features) are in the development stage. The implemented tagset supports the generation of scenes such as Figure 3.

In order to resolve a spatial relation, we find the spatial tags and other features of the FIGURE and GROUND objects that are applicable for the given preposition. For example, if the preposition is *under*, a CANOPY region for the GROUND object is relevant, but not a TOP SURFACE. Various other factors, such as size, must also be considered. With ENCLOSED-IN, the FIGURE must fully fit in the GROUND. For EMBEDDED-IN, only part need fit. For other relations (e.g., NEXT-TO), the objects can be any size, but the FIGURE location might vary. For example, *The mosquito is next to the horse* and *The dog is next to the horse* position the FIGURE in different places, either in the air or on the ground,

Spatial Relation	**Example**	**Partial Conditions**
on-top-surface	*vase on table*	GROUND is UPWARD-SURFACE
on-vertical-surface	*postcard on fridge*	GROUND is VERTICAL-SURFACE
on-downward-surface	*fan on ceiling*	GROUND is DOWNWARD-SURFACE
on-outward-surface	*pimple on nose*	GROUND is SURFACE
pattern/coating-on	*plaid pattern on shirt*	FIGURE is TEXTURE or LAYER
fit-on-custom	*train on track*	SPECIAL BASE PAIRING
ring-on-pole	*bracelet on wrist*	FIGURE=RING-SHAPE, GROUND=POLE-SHAPE
on-vehicle	*man on bus*	GROUND=PUBLIC-TRANSPORTATION
on-region	*on the left side of...*	GROUND=REGION-DESIGNATOR
hang-on	*towel on rod*	FIGURE is HANGABLE
embedded-in	*pole in ground*	GROUND is MASS
embedded-in	*boat in water*	FIGURE is EMBEDDABLE
buried-in	*treasure in ground*	GROUND is TERRAIN
enclosed-in-volume	*bird in cage*	GROUND has ENCLOSURE
enclosed-in-area	*tree in yard*	GROUND is AREA
in-2D-representation	*man in the photo*	GROUND is 2D REPRESENTATION
in-cup	*cherries in bowl*	GROUND has CUP
in-horiz-opening	*in doorway*	GROUND has OPENING
stem-in-cup	*flower in vase*	FIGURE has STEM, GROUND has CUP
wrapped-in	*chicken in the foil*	GROUND is FLEXIBLE/SHEET
member-of-arrangement	*plate in stack*	GROUND is ARRANGEMENT
in-mixture	*dust in air*	FIGURE/GROUND=SUBSTANCE
in-entanglement	*bird in tree*	GROUND has ENTANGLEMENT
fitted-in	*hand in glove*	FIGURE/GROUND=FIT
in-grip	*pencil in hand*	GROUND=GRIPPER

Table 1: Spatial relations for *in* and *on* (approximately half are currently implemented). Similar mappings exist for other prepositions such as *under*, *along*. Handcrafted rules resolve the spatial relation given the object features.

depending on whether the given object is commonly airborne or not. We also note that the FIGURE is normally the smaller object while the GROUND functions as a landmark. So it's normal to say *The flower bed is next to the house*, but unnatural to say **The house is next to the flowerbed*. This is discussed in [9]. See Table 1 for some mappings we make from prepositions to spatial relations.

In order to use the object features described above to resolve the spatial meaning of prepositions, linguistically referenced subregions must also be considered. Spatial relations often express regions relative to an object (e.g., *left side of* in *The chair is on the left side of the room*). The same subregion designation can yield different interpretations, depending on the features of the objects.

- EXTERNAL-VERTICAL-SURFACE: *shutters on the left side of the house*
- INTERIOR-VERTICAL-SURFACE: *picture on the left side of the room*
- REGION-OF-HORIZ-SURFACE: *vase on the left side of the room*
- NEIGHBORING-AREA: *car on the left side of the house*

These regions (when present) are combined with the other constraints on spatial relations to form the final interpretation.

Input text: *A large magenta flower is in a small vase. The vase is under an umbrella. The umbrella is on the right side of a table. A picture of a woman is on the left side of a 16 foot long wall. A brick texture is on the wall. The wall is 2 feet behind the table. A small brown horse is in the ground. It is a foot to the left of the table. A red chicken is in a birdcage. The cage is to the right of the table. A huge apple is on the wall. It is to the left of the picture. A large rug is under the table. A small blue chicken is in a large flower cereal bowl. A pink mouse is on a small chair. The chair is 5 inches to the left of the bowl. The bowl is in front of the table. The red chicken is facing the blue chicken. . .*

Fig. 3: Spatial relations and features: ENCLOSED-IN (*chicken in cage*); EMBEDDED-IN (*horse in ground*); IN-CUP (*chicken in bowl*); ON-TOP-SURFACE (*apple on wall*); ON-VERTICAL-SURFACE (*picture on wall*); PATTERN-ON (*brick texture on wall*); UNDER-CANOPY (*vase under umbrella*); UNDER-BASE (*rug under table*); STEM-IN-CUP (*flower in vase*); LATERALLY-RELATED (*wall behind table*); LENGTH-AXIS (*wall*); DEFAULT SIZE/ORIENTATION (all objects); REGION (*right side of*); DISTANCE (*2 feet behind*); SIZE (*small* and *16 foot long*); ORIENTATION (*facing*).

5 Conclusions and Ongoing and Future Work

In order to represent spatial relations more robustly, much remains to be done at the language, graphical, and application levels.

We are augmenting the system to resolve verbs to semantic frames using information in our SBLR, and mapping those in turn to corresponding poses and spatial relations [5]. Figure 4 illustrates this process, which currently is supported for a limited set of verbs and their arguments. This enhanced capability also requires contextual information about actions and locations that we are acquiring using human annotations obtained via Amazon's Mechanical Turk and by extracting information from corpora using automatic methods [14]. We will be evaluating our software in partnership with a non-profit after-school program in New York City.

Acknowledgments

This work was supported in part by the NSF IIS- 0904361. Any opinions, findings and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsors.

The truck chased the man down the road...

The man ran across the sidewalk...

Fig. 4: Spatial relations derived from verbs. The verbs are mapped to semantic frames which in turn are mapped to VIGNETTES (representing basic contextual situations) given a set of semantic role and values. These, in turn, are mapped to spatial relations. In the first example, the PURSUED (*soldier*) is in a running pose, located on the PATH (*road*), and in front of the PURSUER (truck).

References

1. Adorni, G., Di Manzo, M., Giunchiglia, F.: Natural language driven image generation. COLING pp. 495–500 (1984)
2. Badler, N., Bindiganavale, R., Bourne, J., Palmer, M., Shi, J., Schule, W.: A parameterized action representation for virtual human agents. Workshop on Embodied Conversational Characters, Lake Tahoe (1998)
3. Baker, C., Fillmore, C., Lowe, J.: The Berkeley FrameNet Project. COLING-ACL (1998)
4. Clay, S.R., Wilhelms, J.: Put: Language-based interactive manipulation of objects. IEEE Computer Graphics and Applications pp. 31–39 (1996)
5. Coyne, B., Rambow, O., Hirschberg, J., Sproat, R.: Frame semantics in text-to-scene generation. 14th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (2010)
6. Coyne, B., Sproat, R.: WordsEye: An automatic text-to-scene conversion system. SIGGRAPH, Computer Graphics Proceedings pp. 487–496 (2001)
7. Dupuy, S., Egges, A., Legendre, V., Nugues, P.: Generating a 3d simulation of a car accident from a written description in natural language: The carsim system. Proceedings of ACL Workshop on Temporal and Spatial Information Processing pp. 1–8 (2001)
8. Fellbaum, C.: WordNet: an electronic lexical database. MIT Press (1998)
9. Herskovits, A.: Language and Spatial Cognition: an Interdisciplinary Study of the Prepositions in English. Cambridge University Press, Cambridge, England (1986)
10. Larson, G., Shakespeare, R.: Rendering with Radiance. The Morgan Kaufmann Series in Computer Graphics (1998)
11. Levinson, S.: Space in Language and Cognition: Explorations in Cognitive Diversity. Cambridge University Press, Cambridge (2003)
12. Ma, M.: Automatic Conversion of Natural Language to 3D Animation. Ph.D. thesis, University of Ulster (2006)
13. Simmons, R.: The clowns microworld. Proceedings of TINLAP pp. 17–19 (1998)
14. Sproat, R.: Inferring the environment in a text-to-scene conversion system. First International Conference on Knowledge Capture, Victoria, BC (2001)