Frame Semantics in Text-to-Scene Generation

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

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

Abstract. 3D graphics scenes are difficult to create, requiring users to learn and utilize a series of complex menus, dialog boxes, and often tedious direct manipulation techniques. By giving up some amount of control afforded by such interfaces we have found that users can use natural language to quickly and easily create a wide variety of 3D scenes. Natural language offers an interface that is intuitive and immediately accessible by anyone, without requiring any special skill or training. The WordsEye system (http://www.wordseye.com) has been used by several thousand users on the web to create over 10,000 scenes. The system relies on a large database of 3D models and poses to depict entities and actions. We describe how the current version of the system incorporates the type of lexical and real-world knowledge needed to depict scenes from language.

1 Introduction

The work we describe seeks to bridge the gap between language, graphics, and knowledge by modeling the automatic conversion of text into a new type of semantic representation – a virtual 3D scene. 3D scenes provide an intuitive representation of meaning by making explicit the contextual elements implicit in our mental models. The system we are developing centers on a new type of lexical knowledge representation, which we call a Scenario-Based Lexical Knowledge Resource (SBLR). The SBLR will ultimately include information on the semantic categories of words; the semantic relations between predicates (verbs, nouns, adjectives, and prepositions) and their arguments; the types of arguments different predicates typically take; additional contextual knowledge about the visual scenes various events and activities occur in; and the relationship between this linguistic information and the 3D objects in our objects library. The resulting text-to-scene system, utilizing the SBLR, is applicable to several domains:

- **Education:** Seeing words spring to life makes using language fun and memorable. This suggests many uses in education including ESL, EFL, special needs learning, vocabulary building, and creative storytelling. We have done preliminary testing in some of these areas.

R. Setchi et al. (Eds.): KES 2010, Part IV, LNAI 6279, pp. 375–384, 2010.

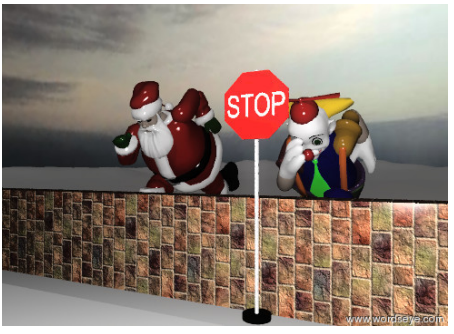
- **Graphics Authoring and Online Communication:** Textually generated scenes can be created in a fraction of the time it takes to make a scene by hand. This increase in speed and lower entry barrier enables a new form of social interaction and promotes "visual banter." We have seen much of this in our online website and gallery.
- **3D Games:** 3D games are painstakingly designed by 3D artists – the malleability of the graphics in games is usually limited to external level design tools and rigid interfaces for character selection and modification. Some recent games, such as Spore (`http://www.spore.com`), allow interesting variability of graphical elements; and spoken language commands are supported by games such as *Tom Clancy's EndWar* [14]. We foresee this trend continuing and games being developed where large parts of the game environment itself are interactively constructed and modified by the players using natural language.

In this paper we describe some of our recent work designing, building, and utilizing the SBLR in order to produce a system with much broader and robust coverage. Examples from the online system are shown in Figure 1. The remainder of the paper is organized as follows. In Section 2 we describe related work with natural language interfaces to 3D graphics. In Section 3 we describe experiences of the system with real online users as well as preliminary testing in schools. In Section 4 we provide an overview of the system. In Section 5 we discuss the SBLR, including the lexicon, semantic relations between lexical items, and frames which group complex relations together. We introduce the notion of a *vignette* and describe the graphical primitives that all semantic representations must resolve to.We conclude and describe future work in Section 6.

2 Related Work

Natural language input has been investigated in a number of very early 3D graphics systems [1][12][4][9] and the Put system [5], which was limited to spatial arrangements of existing objects in a pre-constructed environment. Also, 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 spatial preposition. Several more recent systems target animation rather than scene construction. 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 domain-specific system where short animations were created from natural language descriptions of accident reports. CONFUCIUS [11] is a multimodal animation system that takes as input a single sentence containing an action verb. The system blends animation channels to animate virtual human characters. Another recent system, from the University of Melbourne [16], uses a machine learning-based approach to create animated storyboards on a pre-made virtual stage. In these systems the referenced objects, attributes, and actions are typically relatively small in number or targeted to specific pre-existing domains.

Redundancy by Bob Coyne



Santa Claus is on the white mountain range. He is racing. The clown is 2 feet to the right of santa claus. The clown is racing. A brick wall is 2 feet in front of the clown. The wall is 20 inches tall. It is 20 feet wide. A small stop sign is in front of the wall. It is cloudy.

Infinite Time by Richard Sproat

The clock is one foot in front of the silver wall. The ground has a grass texture. The texture is one foot wide. A silver wall is two feet in front of the clock. A light is fifty feet above the clock.

Fig. 1. Some Examples

As such, these systems have a natural affinity to the SHRDLU system [15] which used natural language to interact with a "robot" living in a closed virtual world.

Our current system is a rewrite and enhancement of the original version of WordsEye [6], which was the first system to use a large library of 3D objects in order to depict scenes in a more general and free-form manner using natural language. The current system contains 2,000 3D objects and 10,000 images and a lexicon of approximately 15,000 nouns. It supports language-based control of spatial relations, textured and colors, collections, and poses; and it handles simple anaphor resolution, allowing for a variety of ways of referring to objects. The earlier WordsEye system handled 200 verbs in a *ad hoc* manner with no systematic semantic modeling of alternations and argument combinations. In the current system, we are instead adding frame semantics to support verbs, event nouns, and stative relations more robustly. The system also does very little inferring of background locations, default poses of characters, and other contextual features – everything must be stated fairly explicitly. As a result, users must describe scenes in somewhat stilted language. But even with these limitations, the process of creating scenes with the system is quick and enjoyable.

3 User Experiences

Earlier versions of the current WordsEye system have been tested online (`http://www.wordseye.com`) over a several year period. A few thousand real-world users have used the system to create and post a large number textually-generated pictures to our online gallery, and in some cases to their personal

webpages and social media sites such as Facebook. The ease and speed of creating and modifying scenes has led to pictures being used as a form of social interaction. We also found, in the course of our testing, that users would try to use the system to do more than it is capable of. In particular they would use language that involved a) complex background settings (e.g., *living room*, *garden*) consisting of many objects b) characters performing actions c) modifications to parts of objects (*the girl's hair is curly*). Our work on the SBLR, to allow the system to better support these areas, has grown out of these experiences.

We also performed some preliminary testing of the system in schools in Spring 2007. After seeing a demo of WordsEye at the Innovate 2007 Exposition (hosted by the State of Virginia Department of Education), K-12 public school teachers from the Albemarle county school system in Virginia asked if they could use it in their classes, believing it to be a useful tool for ESL (English as a second language) remediation, special education, vocabulary enhancement, writing at all levels, technology integration, and art. Feedback from these teachers and their students was quite positive. In one school, with a 10% ESL population, a teacher used it with 5th and 6th graders to reinforce being specific in details of descriptive writing and noted that students are "very eager to use the program and came up with some great pictures." Another teacher tested it with 6th through 8th grade students who were "in a special language class because of their limited reading and writing ability," most reading and writing on a 2nd or 3rd grade level. In addition to its educational value, 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."

4 System Overview

WordsEye consists of multiple components, including a user interface, language processing and 3D graphics. In this paper we focus on the lexical and world knowledge used to convert dependency structures into semantic nodes and roles and the subsequent conversion to graphic frames. The overall system works in the following sequence:

- The user types in text to a webpage.
- The input text is parsed into a dependency tree.
- Anaphora and other coreferences are resolved.
- Lexical items and dependency links are resolved to semantic nodes and roles.
- Semantic relations are converted to graphical relations.
- Default graphically-oriented constraints are inserted (such as putting objects on the ground unless otherwise specified).
- The scene is composed from these constraints and rendered in OpenGL (`http://www.opengl.org`) and optionally ray-traced in the Radiance [10] renderer (`http://radsite.lbl.gov/radiance`).
- The user can provide a title and caption to the finished scene and save it in an online gallery where other users can add comments and create their own pictures in response.

5 SBLR – Lexical, Semantic, and Contextual Information

The SBLR contains lexical, semantic, and contextual information. As such it is related to WordNet and FrameNet, and is, in fact, partially derived from them. We first examine some of the features and limitations of WordNet and FrameNet.

5.1 Wordnet

The WordNet ontology provides a taxonomy of words grouped into separate SYNSETS by word sense and related primarily by HYPERNYM (IS-A) relations. This provides useful information such as the fact that a *chair* is *furniture*. The taxonomy fails, however, to provide a rich set of semantic relations between those entries. For example, the WordNet synset for *princess* is a hyponym of *aristocrat*, but there is no encoding of the basic fact that a princess is also a female. Likewise, WordNet often conflates lexical usage with functional semantic categories. For example, *spoon* is classified as a *container*, which is true in some sense; however, it does not match normal colloquial usage, since a spoon is unlikely to be considered a container by typical speakers of English. Also, while WordNet does encode some part-whole and substance-whole relations, it is missing many very common ones, such as the fact that lamps have lightbulbs and that snowballs are made of snow. In addition, there is no encoding of functional properties of objects such as the fact that a mop is used in cleaning floors. A broader set of semantic relations is crucial to the resolution of lexical references in scene construction.

5.2 Framenet

FrameNet is a digital lexical resource for English that groups related words together into semantic frames [3]. It currently contains 10,000 lexical entries including nouns, verbs, and adjectives. Each lexical unit is associated with one of nearly 800 hierarchically-related semantic frames, where each frame represents the joint meaning of the lexical units in that frame. Each lexical unit is also associated with a set of annotated sentences which map the sentences' constituent parts to their frame-based roles. FrameNet, in total, contains over 135,000 annotated sentences across all lexical units. A FrameNet frame consists of a set of frame-based roles, called *frame elements* (FEs) representing the key roles characterizing the meaning of lexical units in that frame. For example, the COMMERCE_SELL frame includes frame elements for SELLER, GOODS, and BUYER and has lexical units for the verbs *retail*, *sell*, *vend* as well as nouns such as *vendor* and *sale*.

The exact expression of FEs for a given annotated sentence constitutes what FrameNet refers to as a *valence pattern*. Valence patterns are represented as FE and grammatical function (GF) pairs. Grammatical functions are *subject*, *obj*, *second object*, and various other dependent phrases (e.g., *Dep/to*, *Dep/on*, *Dep/with*) which designate the particular prepositional phrase type.

FrameNet provides no semantic information to distinguish the meaning of words in the same frame. For example, the SELF_MOTION frame contains a large number of verbs related only by the fact that the SELF_MOVER moves under its

own power in a directed fashion without a vehicle. As a result, this frame contains strongly related verbs such as *walk* and *stroll* but also verbs with very different manner of motion such as *swim*, and *swing*. Likewise there is no representation in FrameNet of synonymy, antinomy, or other lexical semantic relations.

5.3 SBLR

The SBLR currently contains about 15,000 nouns representing the 3D objects in our library and related words, including a set of semantic relations seeded from FrameNet and augmented as needed. The lexicon was semi-automatically extracted from WordNet, filtering out obscure words and word senses. The resulting lexicon is defined with multiple inheritance to allow the same lexical item to be classified in several ways. In addition, as needed, we encode semantic relations between lexical items via the corresponding SBLR semantic relation. So, to say that a snowball is made of snow, we use the MADE-OF frame.

In addition to their role in verbal adjuncts, prepositions are especially important in text-to-scene generation, since they directly denote spatial relations in a variety of subtle ways [8]. In the SBLR we extend the coverage of frames and lexical units to better model the range of preposition senses. So, for example, *John is on the sidewalk* and *John is on the phone* will constitute different senses of *on* and hence be in different frames. Likewise, *The flower is in the vase*, *The spider is in the web*, and *The goldfish is in the aquarium* represent different, though related, spatial interpretations of *in*.

In the SBLR we also extend FrameNet's notion of valence pattern to directly include semantic and contextual constraints (including SELECTIONAL RESTRICTIONS), drawing upon the semantic types of words and their semantic and contextual relations to other words as defined in the rest of the SBLR. This allows the appropriate frames and frame elements to be assigned to parsed input text. Consider, for example, a few of the semantic interpretations of the preposition *of* and how they are handled by the system in Table 1 and Figure 2.

In order to decompose semantic representations into scenes, we need to supply default instruments, settings, poses, and so on. These choices are captured in the notion of a **vignette**. A vignette is a mapping from a frame-semantic representation to the graphical relations that invoke that scene. For example, the verb *wash* has very different meanings depending on the arguments of the verb. Washing a car takes place outside, often in a driveway with the AGENT

Table 1. SBLR: Semantic Mappings for *of*

Text (A of B)	**Valence Patterns for** *of*	**Resulting Frame Relation**
bowl of cherries	A=container, B=plurality-or-mass	CONTAINER-OF(bowl, cherries)
slab of concrete	A=entity, B=substance	MADE-OF(slab, concrete)
picture of girl	A=representing-entity, B=entity	REPRESENTS(picture, girl)
arm of the chair	A=part-of (B), B=entity	PART-OF (chair, arm)
height of the tree	A=size-property, B=physical-entity	DIMENSION-OF(height, tree)
stack of plates	A=arrangement, B=plurality	GROUPING-OF (stack, plates)

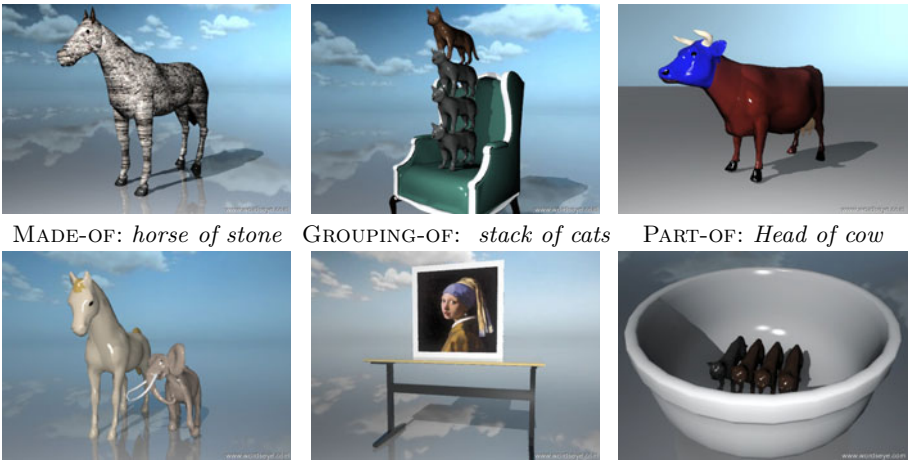


Fig. 2. Depictions of *of*

holding a hose and standing near to the car. Washing dishes usually takes place in a kitchen with the AGENT standing in front of the sink, holding the dishes. Washing the floor likewise invokes yet another prototypical scene. We now consider, in more detail, a simple example: *the truck chased the soldier down the road*.

Valence pattern triggered for *The truck chased the soldier down the road*:

- Frame=COTHEME, FEs=AGENT, COTHEME, PATH, SOURCE, GOAL, ... (This frame contains words that necessarily indicate the motion of two distinct objects.)
- Verb="chase"
- Subject=AGENT, direct-object=COTHEME, dep/down=PATH

Vignette triggered by the above input generates graphical relations putting the soldier (COTHEME) and truck (THEME) on the road (PATH) with the soldier in a running pose in front of the truck:

- Orient-towards: object=THEME, reference=PATH
- Orient-towards: object=COTHEME, reference=PATH
- Position-behind: object=THEME, reference=COTHEME
- Position-on-top: object=THEME, reference=PATH
- Position-on-top: object=COTHEME, reference=PATH
- In-pose: object=COTHEME, pose="running pose"

Figure 3 shows the rendered scene generated using these valence patterns and vignettes.

a) The truck chased the soldier down the road...

b) The soldier ran across the sidewalk...

Fig. 3. Scenes derived from SBLR valence patterns and vignettes. The valence patterns for *chase* and *run* are used to assign semantic frame roles. The semantic relation is then mapped to the appropriate vignettes based on which roles are filled with what. The Vignettes resolve to spatial relations such a) as the COTHEME (*soldier*) being in a running pose and located on the PATH and in front of the THEME (truck) in the case of *chase down* or b) the SELF_MOVER perpendicularly oriented toward the PATH and in a running pose in the case of *run across*. These are combined with explicitly described backgrounds and other objects. In the future, the background settings for different actions will also be defaulted by vignettes.

5.4 Graphical Objects and Relations

Graphical objects and relations are the bedrock to which all semantics must be resolved in order to be depicted. Graphical objects are inserted into the ontology with ISA links referencing existing semantic nodes. This allows them to inherit property values for other objects of their type (such as default size). In addition, almost all 3D objects implicitly contain subtype information (e.g., *dining room chair*, or *antique chinese vase*). And sometimes, 3D objects are compound objects (a lighthouse on a hill) or part of another object (a rose blossom). In all these cases, these properties are represented by semantic relations drawn from our stock of semantic frames.

Graphical objects have various additional geometric and functional properties. These include spatial tags (used in resolving target locations for spatial relations), default size and orientation, and parts. Some of these spatial tags and attributes are described in [6]. The graphical relations which are used in creating a 3D scene include object size, color, position, orientation, texture, aspect ration, facial expressions and poses (for human characters) among others.

The information collected in the SBLR is coming from our own existing resources; external semantic resources such as WordNet, FrameNet, and PropBank, which we are mining for additional information; and information which we will extract from Wikipedia and other corpora . The overall system architecture and the role played by the SBLR is shown in Figure 4.

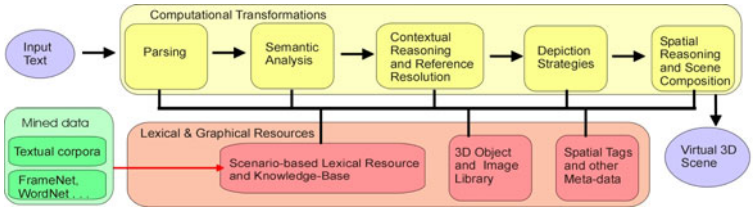


Fig. 4. System Architecture

6 Conclusions and Future Work

The system as described is a work-in-progress. Much remains to be done at the language, graphical, and application levels. Knowledge acquisition, representation, and utilization is our core ongoing task. We are acquiring contextual information such as part relations, default backgrounds for actions, and lexical constraints and verb arguments using Amazon's Mechanical Turk and automatic methods [13]. We have plans in place to evaluate our software in partnership with a non-profit after-school program in New York CIty.

We believe that textually generated scenes offer an exciting new way to interact with a computer and to create visual imagery. It affords access to a wider variety of people than who might otherwise be able to make pictures. And it allows picture-making to be done in completely new settings due to the speed and low overhead involved.

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. We thank Steve Feiner for providing references to some early text-to-scene work and Fadi Biadsy and anonymous reviewers for their useful comments. And we thank the many online WordsEye users for their help in testing and validating this new method of creating 3D scenes and interacting socially with words and pictures.

References

1. Adorni, G., Di Manzo, M., Giunchiglia, F.: Natural language driven image generation. In: 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. In: Workshop on Embodied Conversational Characters, Lake Tahoe (1998)
3. Baker, C., Fillmore, C., Lowe, J.: The Berkeley FrameNet Project. In: COLING-ACL (1998)

4. Boberg, R.: Generating Line Drawings from Abstract Scene Descriptions. Master's thesis, Dept. of Elec. Eng., MIT, Cambridge, MA (1972)
5. Clay, S.R., Wilhelms, J.: Put: Language-based interactive manipulation of objects. In: IEEE Computer Graphics and Applications, pp. 31–39 (1996)
6. Coyne, B., Sproat, R.: WordsEye: An automatic text-to-scene conversion system. In: 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. In: Proceedings of ACL Workshop on Temporal and Spatial Information Processing, pp. 1–8 (2001)
8. Herskovits, A.: Language and Spatial Cognition: an Interdisciplinary Study of the Prepositions in English. Cambridge University Press, Cambridge (1986)
9. Kahn, K.: Creation of Computer Animation from Story Descriptions. Ph.D. thesis, MIT, AI Lab, Cambridge, MA (1979)
10. Larson, G., Shakespeare, R.: Rendering with Radiance. The Morgan Kaufmann Series in Computer Graphics (1998)
11. Ma, M.: Automatic Conversion of Natural Language to 3D Animation. Ph.D. thesis, University of Ulster (2006)
12. Simmons, R.: The clowns microworld. In: Proceedings of TINLAP, pp. 17–19 (1998)
13. Sproat, R.: Inferring the environment in a text-to-scene conversion system. In: First International Conference on Knowledge Capture, Victoria, BC (2001)
14. WikiPedia: Tom Clancy's EndWar, `http://en.wikipedia.org/wiki/Tom_Clancy's_EndWar`
15. Winograd, T.: Understanding Natural Language. Ph.D. thesis, Massachusetts Institute of Technology (1972)
16. Ye, P., Baldwin, T.: Towards automatic animated storyboarding. In: Proceedings of the 23rd National Conference on Artificial Intelligence, vol. 1, pp. 578–583 (2008)