WordsEye: An Automatic Text-to-Scene Conversion System

Presenters: Madhumita Sakthi
Pratyush Kar
Problem Statement

● Creating 3D graphics is a **lengthy** process
  ○ Need to learn complex software package
  ○ Traverse pages of menus
  ○ Tweak parameters, change tools etc.

● Can we use natural language input to automatically convert text into representative 3D scenes?

● **Hard** problem. The system needs to understand:
  ○ Spatial relations (**on the wall** vs **on the table**)
  ○ Actions performed by objects (**riding a bike** vs **riding a horse**)
  ○ Hypernymy (**eg. cat is an animal**) & Hyponymy (**eg. tiger is a cat**)
  ○ etc.
Figure 1: John uses the crossbow. He rides the horse by the store. The store is under the large willow. The small allosaurus is in front of the horse. The dinosaur faces John. A gigantic teacup is in front of the store. The dinosaur is in front of the horse. The gigantic mushroom is in the teacup. The castle is to the right of the store.
WordsEye System Architecture

Lexical, Semantic, Graphical Knowledge-Base

Lexical and morphological analysis
Parse
Reference resolution
Semantic Analysis
Graphical Analysis
Apply graphical constraints
Choose Viewpoint and Render

2D and 3D Objects

Linguistic Analysis

- Text is initially tagged using a POS tagger
  
  ![Example sentence with POS tags]

  John said that the cat was on the table.

- This output is run through a statistical parser like Collin’s parser
Ambiguities like prepositional phrase attachment are resolved
  ○ Based on the statistics of the corpus the parser is trained on (Penn Treebank)

Parse tree is converted to a dependency representation

More convenient for semantic analysis
  ○ For eg. the *large* naughty *black* cat
Handling Anaphoric Expressions

- Matches number & gender features to resolve simple pronominals
  - For eg. *John* uses the crossbow. *He* rides the horse by the store.

- Nouns can also corefer. WordsEye uses WordNet hierarchy to resolve these
  - For eg. *John* said that *the cat* was on the table. *The animal* was next to a lamp.

- Doesn’t handle complex cases like bridging anaphora, etc.
Semantic Analysis

Semantic representation fragments are derived from the dependency structure by semantic interpretation frames

((("node2" (:ENTITY :3D-OBJECTS ("mr_happy")
   :LEXICAL-SOURCE "John" :SOURCE SELF))
("node1" (:ACTION "say" :SUBJECT "node2"
   :DIRECT-OBJECT ("node5" "node4" "node7"...)))
("node5" (:ENTITY :3D-OBJECTS ("cat-vp2842")))
("node4" (:STATIVE-RELATION "on" :FIGURE "node5"
   :GROUND "node7"))
("node7" (:ENTITY :3D-OBJECTS
   ("table-vp14364" "nightstand-vp21374"
   "table-vp4098" "pool_table-vp8359" ...))))
Semantic Analysis (Continued)

- For nouns, WordNet hierarchy is used to look up 3D objects in the database
- Spatial prepositions are handled based on their left and right dependents
  - Spatial relations are object dependent
  - For eg. on the table vs on the wall
- Verbs are handled by semantic frames

```
(SEMANTICS :GENUS say
 :VERB-FRAMES
  ((VERB-FRAME
    :NAME SAY-BELIEVE-THAT-S-FRAME
    :REQUIRED (SUBJECT THAT-S-OBJECT)
    :OPTIONAL (ACTIONLOCATION ACTIONTIME))
```
Spatial Relations

- Defines the basic layout of the scenes
- Mostly used prepositions are on, under, beyond, etc
- For The bird is on the cat
- Find a top surface tag for the cat (on its back)
- A base tag for the bird (under its feet).
- Then reposition the bird so that its feet are on the cat’s back
Spatial Relations

- Spatial tags for stems are applied to any object with a long, thin base leading to a thicker or wider top area.
- Some objects with stems are stop signs, umbrellas, palm trees and street lamps.
Poses and Grips

● Standalone poses consist of a character in a particular body position
● Such as waving, running, bowing, kneeling
● Specialized usage poses involve a character using a specific instrument or vehicle
● Generic usage pose: character interacting with a generic stand-in object

Figure 9: Usage pose for a 10-speed.

Figure 10: “Throw small object” pose and “hold wine bottle” grip.
Inverse Kinematics

- Characters interact directly with the environment using IK.
- Pointing towards a desired direction using IK.
- The character is first put behind the object in the push large object pose.
- The hands are moved using IK to the handle or vertical surface of the object.
- This technique relies on object tags for handle or vertical surface in order to determine the target position for the IK.

Figure 12: Spatial tag for “push handle” of baby carriage, indicated by the box around the handle.

Figure 13: The lawn mower is 5 feet tall. John pushes the lawn mower. The cat is 5 feet behind John. The cat is 10 feet tall.
Attributes

- Color and transparency are applied to the object as **surface attributes**
- The shape of the object can be modified using shape displacements in the Mirai animation system
- For example, in a human face, there can be separate displacements for a smile and a wink
- Size (e.g., large, small) and aspect ratio (e.g., flattened, squashed) are controlled by manipulating the 3D object’s **transform matrix**
The Depiction Process

- Set of low-level depictors are 3D objects with their spatial and graphical properties

  Semantic representation → Interpret for ‘?’ → Apply depictor → Resolve conflicts

- Read in referenced 3D models
- Build up the scene with each depictor
- Adjust camera and finally, render
Depiction Rules

Kick

- **Case1**: no PATH or RECIPIENT, size greater than 3 feet
- Position: ACTOR directly behind DIRECT-OBJECT
- **Case2**: size less than 3 feet, Pose: kick object
- **Case3**: PATH and RECIPIENT
  - ACTOR Path: DIRECT-OBJECT between ACTOR’s foot and RECIPIENT
  - Orientation: ACTOR facing RECIPIENT
  - Pose: catch, RECIPIENT [tentative]
  - Orientation: RECIPIENT facing ACTOR [tentative]
Depiction Rules & Implicit constraints

- Spinning
- Spatial-Relation: above, SUBJECT, circular arrow 3D model
- DEPICTION-RULE ENTITY cowboy
- Pose: wear cowboy hat, ACTOR

Implicit constraints:

- The lamp is on the table. The glass is next to the lamp
- “If X is next to Y, X is not already on a surface, and X is not an airborne object (e.g., a helium balloon)” then “Put X on the same surface as Y”.
Conflicting constraints

1. POSE: John in kick pose
2. PATH: Ball between John’s foot and Mary
3. ORIENTATION: John facing Mary
4. POSE: Mary in catch pose [tentative]
5. ORIENTATION: Mary facing John [tentative]
6. POSITION: John 10 feet from Mary in Z axis [tentative] / John 20 feet from Mary in Z axis
7. POSITION: John 0 feet from Mary in X axis [tentative] / John 30 feet from Mary in X axis
Interpretations, Activities, Environment

- Interpret a vague text like “John went to the store” and add details
- “John rode to the store” depicted as riding a bicycle
- Look for objects whose functional properties are compatible with the instrument type demanded by the verb
- “The flower is blue”, supply a background

- WordsEye does not have enough real-world knowledge to represent “John filled his car with gas”
WordsEye

First Grade homework

Bee sat on a duck.
Duck sat on a hen.
Hen sat on a pig.
Pig sat on a cow.
Cow sat on a sheep.

WordsEye interpretation
Conclusion

- They have claimed that WordsEye represents a new approach to creating 3D scenes and images.
- Also, the intention is not to replace traditional 3D tools but to augment them by quickly creating a scene and refining it later.
- Since semantic intent can be ambiguous, it leads to various interesting results.
- Adjust language to control a depiction more precisely.

Figure 17: The devil is in the details.
Limitations of the Solution

- Most of the methods mentioned in this paper are hand coded.
- Adding new 3D object models or words senses is largely a manual process.
- Very hard to automatically evaluate the system.
Research Directions

• Sometimes representing text in a 3D scene requires adding information that was not explicitly stated
  ○ Eg. *John went to the store.*

• The system can automatically formulate clarification questions to construct the scene