Text to 3D Scene Generation with Rich Lexical Grounding

Daniel Crockett
Task

Generate appropriate 3D scene based on natural language description

"There is a desk and there is a notepad on the desk. There is a pen next to the notepad."
Motivation

- Creating 3D scenes difficult and time consuming
- Much easier to describe a scene and have it generated
Living room with brick walls, large, factory style windows on one wall and part of the roof, a wood floor, two couches arranged in an L-shape with a small coffee table on top of a rug. Succulent plants on the window sill and a unique chandelier.
Motivation

- Robotics (visual representation of commands)
- Video Game Design
- Architecture
Data Set

There is a chair and a table
Data Set

- There is a chair and a circular table in the middle of a floral print room.
- A corner widow room with a table and chair sitting to the east side.
- There’s a dresser in the corner of the room, and a yellow table with a brown wooden chair.
Learned Model

Train model to discriminate between described scene and 4 random scenes

“There is a desk and there is a notepad on the desk. There is a pen next to the notepad.”
Learned Model

Uses binary-valued features of unigram/bigram co-occurrence with object category or model ID.

(desk, modelId:132)

(the notepad, category:notepad)
Learned Model

- 0.715 accuracy using model ID only
- 0.833 accuracy using object categories also
- Co-occurrence weights used to predict model for generation
Rule Based Model

Input Text

“There is a room with a table and a cake. There is a red chair to the right of the table.”

a) Scene Template

b) Geometric Scene
c) 3D Scene
Rule Based Model

Scene Template Parsing

- Parse sentences with Stanford CoreNLP pipeline
- Resolve coreference with Stanford coreference system
Rule Based Model

Scene Template Parsing

- Identify objects using head word of NP as category
- Identify attributes using NP and ADJ
- Dependency patterns used to extract spatial relations
- Predefined and learned spatial relations
Rule Based Model

Infer Implicit Objects and Relations in Scene:

Learn priors learned from data

\[ P_{surf}(S_n|C_c) = \frac{\text{count}(C_c \text{ on surface with } S_n)}{\text{count}(C_c)} \]
Rule Based Model

Select 3D models from database and arrange according to understood relations

- **Input Text**
  - "There is a desk and a keyboard and a monitor."

- **Basic**

- **+Support Hierarchy**

- **+Relative Positions**

- **No Relations**

- **Predefined Relations**

- **Learned Relations**

- "There is a coffee table and there is a lamp behind the coffee table. There is a chair in front of the coffee table."
Combined Model

Learned model: Select 3D models

Rule based model: identify potential objects, arrange the 3D scene

Score a 3D model $m$ based on description $d$:

$$m = \arg \max_{m \in \{c\}} \lambda_d \sum_{\phi_i \in \phi(d)} \theta_{(i,m)} + \lambda_x \sum_{\phi_i \in \phi(x)} \theta_{(i,m)}$$
Learned Lexical Groundings

<table>
<thead>
<tr>
<th>text</th>
<th>category</th>
<th>text</th>
<th>category</th>
</tr>
</thead>
<tbody>
<tr>
<td>chair</td>
<td>Chair</td>
<td>round</td>
<td>RoundTable</td>
</tr>
<tr>
<td>lamp</td>
<td>Lamp</td>
<td>laptop</td>
<td>Laptop</td>
</tr>
<tr>
<td>couch</td>
<td>Couch</td>
<td>fruit</td>
<td>Bowl</td>
</tr>
<tr>
<td>vase</td>
<td>Vase</td>
<td>round table</td>
<td>RoundTable</td>
</tr>
<tr>
<td>sofa</td>
<td>Couch</td>
<td>laptop</td>
<td>Computer</td>
</tr>
<tr>
<td>bed</td>
<td>Bed</td>
<td>bookshelf</td>
<td>Bookcase</td>
</tr>
</tbody>
</table>

red cup  round yellow table  green room  black top  tan love seat  black bed  open window
Results

Humans asked to rate how well generated 3D scene matched description on scale 1-7:

<table>
<thead>
<tr>
<th>method</th>
<th>Seeds</th>
<th>MTurk</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>2.03 (1.88 – 2.18)</td>
<td>1.68 (1.57 – 1.79)</td>
</tr>
<tr>
<td>learned</td>
<td>3.51 (3.23 – 3.77)</td>
<td>2.61 (2.40 – 2.84)</td>
</tr>
<tr>
<td>rule</td>
<td>5.44 (5.26 – 5.61)</td>
<td>3.15 (2.91 – 3.40)</td>
</tr>
<tr>
<td>combo</td>
<td>5.23 (4.96 – 5.44)</td>
<td>3.73 (3.48 – 3.95)</td>
</tr>
<tr>
<td>human</td>
<td>6.06 (5.90 – 6.19)</td>
<td>5.87 (5.74 – 6.00)</td>
</tr>
</tbody>
</table>
## Results

<table>
<thead>
<tr>
<th>Description</th>
<th>random</th>
<th>learned</th>
<th>rule</th>
<th>combo</th>
</tr>
</thead>
</table>
| Seed sentence:  
*There is a desk and a computer.* | ![Random Image] | ![Learned Image] | ![Rule Image] | ![Combo Image] |
| MTurk sentences:  
*A round table is in the center of the room with four chairs around the table. There is a double window facing west. A door is on the east side of the room.* | ![Image] | ![Image] | ![Image] | ![Image] |
| *In between the doors and the window, there is a black couch with red cushions, two white pillows, and one black pillow. In front of the couch, there is a wooden coffee table with a glass top and two newspapers. Next to the table, facing the couch, is a wooden folding chair.* | ![Image] | ![Image] | ![Image] | ![Image] |
Results

Scene Similarity Metric:

\[
\text{ASTS}(s, z) = \max_A \frac{\sum_{(n, n') \in A} S(n, n')} {J(A) + |A|}.
\]

\(J(A)\)- sum of number of unaligned nodes in scene template and number of unaligned objects is scene

\(S(n, n') = \{1 \text{ if model id matches, } 0.5 \text{ if category matches, } 0 \ o.t.\} \)

<table>
<thead>
<tr>
<th>method</th>
<th>Human</th>
<th>ASTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>1.68</td>
<td>0.08</td>
</tr>
<tr>
<td>learned</td>
<td>2.61</td>
<td>0.23</td>
</tr>
<tr>
<td>rule</td>
<td>3.15</td>
<td>0.32</td>
</tr>
<tr>
<td>combo</td>
<td>3.73</td>
<td>0.44</td>
</tr>
</tbody>
</table>
Scene Generation Errors

Wood table and four wood chairs in the center of the room.

There is a black and brown desk with a table lamp and flowers.

There is a white desk, a black chair, and a lamp in the corner of

There in the middle is a table, on the table is a cup.
Critique

Learned Model:

- Paper ambiguous on how unigrams and bigrams are selected for features (every unigram/bigram in sentence?)
- No explanation of why they chose to only use unigram/bigrams
- Model exists for grounding objects and attributes to certain 3D models, very inflexible for other applications (no actual knowledge of color, material, etc. of 3D models).
Critique

Learned Model:

- Select four objects with highest score: Why not use a score threshold?

- Unconvinced that retaining spelling and grammatical errors will be useful based on small size of the dataset.
Critique (Pros)

Use of human evaluation was important to evaluation of model
  - Very hard to have an automated way of evaluation

Tested to see if proposed automatic metric was correlated with human evaluation

Focus on statistical significance in results
Future Work

- Creating a model that generalizes to 3D models unseen in testing
  - Use of CNN on 3D model to extract features
- Actual 3D model generation based on text descriptions
- Incorporation of world knowledge into scene generation
Questions