Text to 3D Scene Generation with Rich Lexical Grounding

Angel Chang     Will Monroe     Manolis Savva
Christopher Potts     Christoper D. Manning
Stanford University

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

ACL-IJCNLP     July 27, 2015     Beijing, China
Outline

- Introduction and prior work
- Dataset
- Lexical learning
- Generation with lexical grounding
- Evaluation
- Challenges and Conclusion
Outline

• Introduction and prior work
  • Dataset
  • Lexical learning
• Generation with lexical grounding
• Evaluation
• Challenges and conclusion
The art of 3D scene design
The art of 3D scene design

Call of Duty: Advanced Warfare
[Activision / Sledgehammer Games]
The art of 3D scene design

Call of Duty: Advanced Warfare
Activision / Sledgehammer Games

Toy Story 3
Disney / Pixar
The art of 3D scene design

Call of Duty: Advanced Warfare
[Activision / Sledgehammer Games]

Toy Story 3
[Disney / Pixar]

“Modern: Plywood, Plastic & Polished Metal”
[Homedit Interior Design & Architecture]
Generating 3D scenes from text
TOYS’ POV -- An idyllic day care classroom, filled with the happy bustle of four- and five-year-olds, playing with toys -- dinosaurs, a baby doll, a pink Teddy bear, a Ken doll. ...

A Tonka Truck races forward, then backs up in a quick 180 arc, revealing a large pink Teddy bear, LOTSO, in its bed. Lotso taps a Tinker Toy cane and the truck bed rises, “dumping” him out. Like Bob Hope stepping off the links in Palm Springs, Lotso exudes an easy, cheerful charisma.

(Screenplay by Michael Arndt)
Selected prior work

SHRDLU (Winograd, 1972)  WordsEye (Coyne and Sproat, 2001)
Scene generation pipeline

There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.

(Chang et al., 2014)
Scene generation pipeline

There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.

(Chang et al., 2014)
Scene generation pipeline

There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.

(Chang et al., 2014)
Scene generation pipeline

There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.

(Chang et al., 2014)
Handling lexical variety

- sofa
- couch
- loveseat
- dresser
- chest of drawers
- cabinet
Identifying object mentions

Wood table and four wood chairs in the center of the room
Identifying object mentions

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

Can we fix this by learning from data?
Outline

• Introduction and prior work
• Dataset
• Lexical learning
• Generation with lexical grounding
• Evaluation
• Challenges and conclusion
Outline

- Introduction and prior work
- Dataset
- Lexical learning
- Generation with lexical grounding
- Evaluation
- Challenges and conclusion
Dataset

There is a bed and there is a chair next to the bed.
There is a bed and there is a chair next to the bed.
Structure of a 3D scene
Structure of a 3D scene

```json
{
    'modelID': '7bdc0aac',
    'position': [118.545639, 97.979499, 3.098599],
    'scale': 0.087807,
    'rotation': -1.088704
}
```
Structure of a 3D scene

```
{
  'modelID': '7bdc0aacc',
  'position': [118.545639, 97.979499, 3.098599],
  'scale': 0.087807,
  'rotation': -1.088704
}
```

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>ellington armchair</td>
</tr>
<tr>
<td>id</td>
<td>7bdc0aacc</td>
</tr>
<tr>
<td>tags</td>
<td>armchair, chair, ellington, haughton, sam, seating, woodmark</td>
</tr>
<tr>
<td>category</td>
<td>Chair</td>
</tr>
<tr>
<td>wnlemmas</td>
<td>armchair</td>
</tr>
<tr>
<td>unit</td>
<td>0.028974</td>
</tr>
<tr>
<td>up</td>
<td>[0, 0, 1]</td>
</tr>
<tr>
<td>front</td>
<td>[0, -1, 0]</td>
</tr>
</tbody>
</table>
Structure of a 3D scene

```json
{
    'modelID': '7bdc0aac',
    'position': [118.545639, 97.979499, 3.098599],
    'scale': 0.087807,
    'rotation': -1.088704
}
```

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td>ellington armchair</td>
</tr>
<tr>
<td>id</td>
<td>7bdc0aac</td>
</tr>
<tr>
<td>tags</td>
<td>armchair, chair, ellington, haughton, sam, seating, woodmark</td>
</tr>
<tr>
<td>category</td>
<td>Chair</td>
</tr>
<tr>
<td>wnlemmas</td>
<td>armchair</td>
</tr>
<tr>
<td>unit</td>
<td>0.028974</td>
</tr>
<tr>
<td>up</td>
<td>[0, 0, 1]</td>
</tr>
<tr>
<td>front</td>
<td>[0, -1, 0]</td>
</tr>
</tbody>
</table>
There is a bed and there is a chair next to the bed.
There is a bed and there is a chair next to the bed.

The room has three windows on one wall. There is a red bed in the back of the room. Along side the bed is a side chair that is red and white.

This room has a bed with red bedding against the wall. Next to the bed is a chair.

there is a antique looking bed with red covers and pillows in a room. next to it is a recliner chair with red padding. also there are windows.

there is a bed with five pillows on it, and next to it is a chair

There is a bed in the room with two pillows and a small chair near to the right side of it.

There is a large grey bed in the bottom right corner of the room. Above the bed is a small black chair.

Floor to ceiling windows on back wall. Green bed with two pillows and black blanket. Lights recessed into right side wall. Light wood flooring. A chair is in the upper right hand corner

There is a bed on the side of the room. There is a chair in the corner, next to the windows.

I see a bed and a chair.
Dataset

There is a bed and there is a chair next to the bed.

The room has three windows on one wall. There is a red bed in the back of the room. Alongside the bed is a side chair that is red and white.

This room has a bed with red bedding against the wall. Next to the bed is a chair.

There is a bed in the back of the room. Alongside the bed is a side chair that is red and white.

There is a large grey bed in the bottom right corner of the room. Above the bed is a small black chair.

Floor to ceiling windows on back wall. Green bed with two pillows and black blanket. Lights recessed into right side wall. Light wood flooring. A chair is in the upper right hand corner.

There is a bed on the side of the room. There is a chair in the corner next to the windows.

I see a bed and a chair.
There is a bed and there is a chair next to the bed.

There is a bed and there is a chair next to the bed.

There is a bed and there is a chair next to the bed.

Floor to ceiling windows on back wall. Green bed with two pillows and black blanket. Lights recessed into right side wall. Light wood flooring. A chair is in the upper right hand corner.

There is a bed on the side of the room. There is a chair in the corner, next to the windows.

I see a bed and a chair.
There is a bed and there is a chair next to the bed.

The room has three windows on one wall. There is a red bed in the back of the room. Along side the bed is a side chair that is red and white.

This room has a bed with red bedding against the wall. Next to the bed is a chair.

There is a bed with red bedding against the wall. Next to the bed is a chair.

There is an antique looking bed with red covers and pillows in a room. Next to it is a recliner chair with red padding. Also there are windows.

There is a bed with five pillows on it, and next to it is a chair.

There is a large grey bed in the bottom right corner of the room. Above the bed is a small black chair.

60 seed sentences 1128 scenes 4284 scene descriptions
Outline

• Introduction and prior work
• Dataset
• Lexical learning
• Generation with lexical grounding
• Evaluation
• Challenges and conclusion
Discrimination task

*brown room with a refrigerator in the back corner*
Discrimination task

brown room with a refrigerator in the back corner
Learning lexical items

- One-vs.-all logistic regression
- Features: $\{\text{language, object}\}$
  - language: bag-of-words / bag-of-bigrams
  - object: model id / category

- brown
- brown room
- room
- room with
- with
- ...
- room01
- room02
- 7bdc0aac
- cat:Room
- cat:Refrigerator
- ...

Discrimination results

- Accuracy (% correct scenes identified)

<table>
<thead>
<tr>
<th></th>
<th>Random set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model ids only</td>
<td>71.5%</td>
</tr>
<tr>
<td>Model ids + categories</td>
<td><strong>83.3%</strong></td>
</tr>
</tbody>
</table>
### Lexical grounding examples

<table>
<thead>
<tr>
<th>text</th>
<th>category</th>
</tr>
</thead>
<tbody>
<tr>
<td>chair</td>
<td>Chair</td>
</tr>
<tr>
<td>couch</td>
<td>Couch</td>
</tr>
<tr>
<td>sofa</td>
<td>Couch</td>
</tr>
<tr>
<td>fruit</td>
<td>Bowl</td>
</tr>
<tr>
<td>bookshelf</td>
<td>Bookcase</td>
</tr>
</tbody>
</table>
Lexical grounding examples

red cup  round yellow table  green room  black top

tan love seat  black bed  open window
Outline

• Introduction and prior work
• Dataset
• Lexical learning
• Generation with lexical grounding
• Evaluation
• Challenges and conclusion
There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.
There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.
Baseline

There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.
Baseline

There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.
Baseline

There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.
Baseline

There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.

choose top $k$  
($k = 4$)

$K = 4$, average number of objects in human-constructed scenes
Baseline

There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.

choose top $k$  
$(k = 4)$

No relationship enforced between objects! Combine with rule-based parser?
Rule-based parsing

There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.
Rule-based parsing

There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.

- Identify **object categories using noun phrases**

(Chang et al., 2014)
There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.

- Identify object categories using noun phrases
- Identify attributes and keywords using modifiers and dependency patterns

(Chang et al., 2014)
There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.

- Identify object categories using noun phrases
- Identify attributes and keywords using modifiers and dependency patterns
- Identify spatial relations using dependency patterns

(Chang et al., 2014)
There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.

- Identify object categories using noun phrases
- Identify attributes and keywords using modifiers and dependency patterns
- Identify spatial relations using dependency patterns
- Look up objects from DB using categories and keywords

(Chang et al., 2014)
there is a room with a wooden desk and a black lamp
there is a room with a wooden desk and a **black lamp**

\[ c = \text{argmax}_c \sum_{\varphi_i \in \varphi(p)} \theta(i, c) \]
there is a room with a wooden desk and a **black lamp**

\[
c = \arg\max_c \sum_{\phi_i \in \phi(p)} \theta_{(i,c)}
\]

<table>
<thead>
<tr>
<th>Object</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lamp</td>
<td>2.304</td>
</tr>
<tr>
<td>Table</td>
<td>0.622</td>
</tr>
<tr>
<td>Vase</td>
<td>-0.310</td>
</tr>
</tbody>
</table>
there is a room with a wooden desk and a black lamp

\[ c = \arg\max_c \sum_{\phi_i \in \phi(p)} \theta_{(i,c)} \]

\[ m = \arg\max_{m \in c} \left( \lambda_d \sum_{\phi_i \in \phi(d)} \theta_{(i,m)} + \lambda_x \sum_{\phi_i \in \phi(x)} \theta_{(i,m)} \right) \]

Lamp 2.304
Table 0.622
Vase -0.310
there is a room with a wooden desk and a black lamp

$Lamp = \text{argmax} \sum_{i \in \Phi(p)} \theta(i, c)$

$m = \text{argmax} \left( \lambda_d \sum_{i \in \Phi(d)} \theta(i, m) + \lambda_x \sum_{i \in \Phi(x)} \theta(i, m) \right)$
there is a room with a wooden desk and a black lamp

\[
c = \arg \max_c \sum_{\varphi_i \in \varphi(p)} \theta_{(i,c)}
\]

\[
m = \arg \max_{m \in c} \left( \lambda_d \sum_{\varphi_i \in \varphi(d)} \theta_{(i,m)} + \lambda_x \sum_{\varphi_i \in \varphi(x)} \theta_{(i,m)} \right)
\]

Lamp  2.304
Table  0.622
Vase   -0.310

0.302  0.460  -0.021
There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.
Scene generation pipeline

There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.

(Chang et al., 2014)
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.
Outline

- Introduction and prior work
- Dataset
- Lexical learning
- Generation with lexical grounding
- Evaluation
- Challenges and conclusion
Evaluation

- Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)
Evaluation

- Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)

- Compare scenes generated with four methods against human-built scenes
Evaluation

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.
Evaluation

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.
Evaluation

- Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)

- Compare scenes generated with 4 methods (random, lexical baseline, rule-based-parser, combined) against human-built scenes
Evaluation

• Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)

• Compare scenes generated with 4 methods (random, lexical baseline, rule-based-parser, combined) against human-built scenes

• Two sets of scene descriptions
  Seed: seed sentences
  Mturk: descriptions provided by turkers
There is a bed and there is a chair next to the bed.
Dataset

Seed

There is a bed and there is a chair next to the bed.

Simple, no modifiers
There is a bed and there is a chair next to the bed.
There is a bed and there is a chair next to the bed.

There is a bed in the room with two pillows and a small chair near to the right side of it.

There is a large grey bed in the bottom right corner of the room. Above the bed is a small black chair.

There is a bed in the back of the room. Along side the bed is a side chair that is red and white.

This room has a bed with red bedding against the wall. Next to the bed is a chair.

There is a bed with five pillows on it and next to it is a chair.

Floor to ceiling windows on back wall. Green bed with two pillows and black blanket. Lights recessed into right side wall. Light wood flooring. A chair is in the upper right hand corner.

There is a bed on the side of the room. There is a chair in the corner, next to the windows.

I see a bed and a chair.
Dataset

There is a bed and there is a chair next to the bed.

Seed

There is a bed and there is a chair next to the bed.

More complex, varied language

Mturk

The room has three windows on one wall. There is a red bed in the back of the room. Alongside the bed is a side chair that is red and white.

This room has a bed with red bedding against the wall. Next to the bed is a chair.

there is a antique looking bed with red covers and pillows in a room. next to it is a recliner chair with red padding. also there are windows.

The room has three windows on one wall. There is a red bed in the back of the room. Alongside the bed is a side chair that is red and white.

There is a bed in the room with two pillows and a small chair near to the right side of it.

Floor to ceiling windows on back wall. Green bed with two pillows and black blanket. Lights recessed into right side wall. Light wood flooring. A chair is in the upper right hand corner.

There is a large grey bed in the bottom right corner of the room. Above the bed is a small black chair.

There is a bed on the side of the room. There is a chair in the corner, next to the windows.

I see a bed and a chair.
Evaluation Results

Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)

Method

Random
Lexical baseline
Rule-based parser
Combined
Human-built

168 participants, average 4.2 ratings per scene-description pair
Evaluation Results

Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)

<table>
<thead>
<tr>
<th>Method</th>
<th>Seed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>2.03</td>
</tr>
<tr>
<td>Lexical baseline</td>
<td>3.51</td>
</tr>
<tr>
<td>Rule-based parser</td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td></td>
</tr>
<tr>
<td>Human-built</td>
<td></td>
</tr>
</tbody>
</table>

168 participants, average 4.2 ratings per scene-description pair
Evaluation Results

Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)

<table>
<thead>
<tr>
<th>Method</th>
<th>Seed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>2.03</td>
</tr>
<tr>
<td>Lexical baseline</td>
<td>3.51</td>
</tr>
<tr>
<td>Rule-based parser</td>
<td></td>
</tr>
<tr>
<td>Combined</td>
<td></td>
</tr>
<tr>
<td>Human-built</td>
<td></td>
</tr>
</tbody>
</table>

168 participants, average 4.2 ratings per scene-description pair
Evaluation Results

Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)

<table>
<thead>
<tr>
<th>Method</th>
<th>Seed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>2.03</td>
</tr>
<tr>
<td>Lexical baseline</td>
<td>3.51</td>
</tr>
<tr>
<td>Rule-based parser</td>
<td>5.44</td>
</tr>
<tr>
<td>Combined</td>
<td></td>
</tr>
<tr>
<td>Human-built</td>
<td>6.06</td>
</tr>
</tbody>
</table>

168 participants, average 4.2 ratings per scene-description pair
# Evaluation Results

Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)

<table>
<thead>
<tr>
<th>Method</th>
<th>Seed</th>
<th>Mturk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>2.03</td>
<td>1.68</td>
</tr>
<tr>
<td>Lexical baseline</td>
<td>3.51</td>
<td>2.61</td>
</tr>
<tr>
<td>Rule-based parser</td>
<td>5.44</td>
<td>3.15</td>
</tr>
<tr>
<td>Combined</td>
<td>6.06</td>
<td>5.87</td>
</tr>
</tbody>
</table>

168 participants, average 4.2 ratings per scene-description pair
**Evaluation Results**

Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)

<table>
<thead>
<tr>
<th>Method</th>
<th>Seed</th>
<th>Mturk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>2.03</td>
<td>1.68</td>
</tr>
<tr>
<td>Lexical baseline</td>
<td>3.51</td>
<td>2.61</td>
</tr>
<tr>
<td>Rule-based parser</td>
<td>5.44</td>
<td>3.15</td>
</tr>
<tr>
<td>Combined</td>
<td>6.06</td>
<td>5.87</td>
</tr>
<tr>
<td>Human-built</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

168 participants, average 4.2 ratings per scene-description pair
Evaluation Results

Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)

<table>
<thead>
<tr>
<th>Method</th>
<th>Seed</th>
<th>Mturk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>2.03</td>
<td>1.68</td>
</tr>
<tr>
<td>Lexical baseline</td>
<td>3.51</td>
<td>2.61</td>
</tr>
<tr>
<td>Rule-based parser</td>
<td>5.44</td>
<td>3.15</td>
</tr>
<tr>
<td>Combined</td>
<td>5.23</td>
<td>3.73</td>
</tr>
<tr>
<td>Human-built</td>
<td>6.06</td>
<td>5.87</td>
</tr>
</tbody>
</table>

168 participants, average 4.2 ratings per scene-description pair
Evaluation Results

Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)

<table>
<thead>
<tr>
<th>Method</th>
<th>Seed</th>
<th>Mturk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>2.03</td>
<td>1.68</td>
</tr>
<tr>
<td>Lexical baseline</td>
<td>3.51</td>
<td>2.61</td>
</tr>
<tr>
<td>Rule-based parser</td>
<td>5.44</td>
<td>3.15</td>
</tr>
<tr>
<td>Combined</td>
<td>5.23</td>
<td>3.73</td>
</tr>
<tr>
<td>Human-built</td>
<td>6.06</td>
<td>5.87</td>
</tr>
</tbody>
</table>

168 participants, average 4.2 ratings per scene-description pair
Outline

• Introduction and prior work
• Dataset
• Lexical learning
• Generation with lexical grounding
• Evaluation
• Challenges and conclusion
## Evaluation Results

Turkers rated fidelity of generated scenes on a scale of 1 (poor) to 7 (good)

<table>
<thead>
<tr>
<th>Method</th>
<th>Seed</th>
<th>Mturk</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>2.03</td>
<td>1.68</td>
</tr>
<tr>
<td>Lexical baseline</td>
<td>3.51</td>
<td>2.61</td>
</tr>
<tr>
<td>Rule-based parser</td>
<td>5.44</td>
<td>3.15</td>
</tr>
<tr>
<td>Combined</td>
<td>5.23</td>
<td>3.73</td>
</tr>
<tr>
<td>Human-built</td>
<td>6.06</td>
<td>5.87</td>
</tr>
</tbody>
</table>

168 participants, average 4.2 ratings per scene-description pair
Generated scene examples

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.
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.
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.
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.
Generated scene examples

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.
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.
Remaining Challenges

● Grounding of spatial relations

There in the middle is a \textit{table}. Facing the couch

● Coreference

\textit{There in the middle is a table. On the table is a cup.}
Summary

- Learning of lexical grounding to handle linguistic variation in scene description

red cup  round yellow table
Summary

- Learning of lexical grounding to handle linguistic variation in scene description
- Combined rule-based parser and learned lexical groundings for scene generation
Summary

- Learning of lexical grounding to handle linguistic variation in scene description
- Combined rule-based parser and learned lexical groundings for scene generation
- Evaluation demonstrating improved text to scene generation
Thank you!

Dataset is publicly available
http://nlp.stanford.edu/data/text2scene.shtml