

Understanding Scene Descriptions for Text to 3D Scene Generation



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

Will Monroe
CS 224U: Natural Language Understanding
May 11, 2015

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



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)



Help



Meta



Undo



Redo



Copy



Paste



Delete



Tumble



Save



Close

What's in a 3D scene



Model Search

Search

chair



chair



chair



chair



chair



chair



chair



chair



chair



chair



school ...



desk ch...



dining c

comput



Help



Meta



Undo



Redo



Copy



Paste



Delete



Tumble



Save



Close

What's in a 3D scene

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

Model Search

Search

chair



chair



chair



chair



chair



chair



chair



chair



chair



chair



school ...



desk ch...



dining c

comput

What's in a 3D scene

```
{  
  'modelID': '7bdc0aac'  
  'position': [118.545639,  
               97.9  
               3.09  
  'scale': 0.08780  
  'rotation': -1.08  
}
```



Field	Value
name	ellington armchair
id	7bdc0aac
tags	armchair, chair, ellington, haughton, sam, seating, woodmark
category	Chair
wnlemmas	armchair
unit	0.028974
up	[0, 0, 1]
front	[0, -1, 0]

What's in a 3D scene

```
{  
  'modelID': '7bdc0aac'  
  'position': [118.545639,  
              97.9  
              3.09  
  'scale': 0.08780  
  'rotation': -1.08  
}
```

human-tagged
keywords &
categories



WordNet!

size & orientation
suggestions

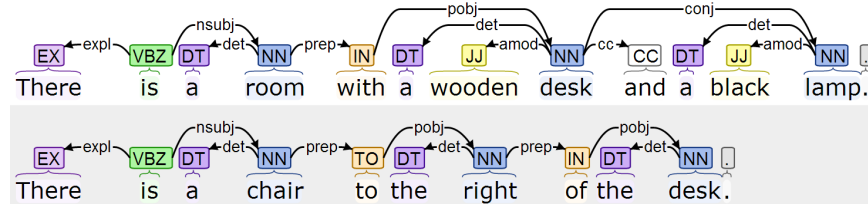
Field	Value
name	ellington armchair
id	7bdc0aac
tags	armchair, chair, ellington, haughton, sam, seating, woodmark
category	Chair
wnlemmas	armchair
unit	0.028974
up	[0, 0, 1]
front	[0, -1, 0]

Scene Generation Pipeline

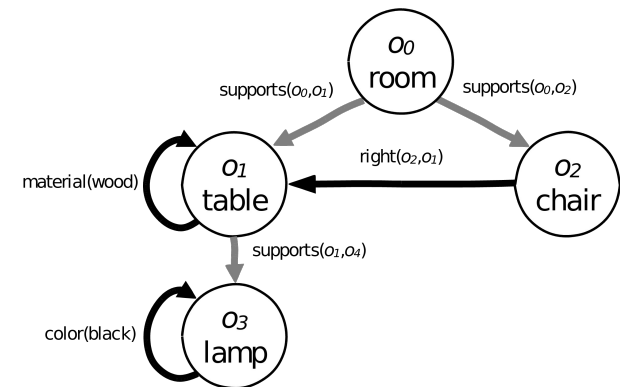
There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.

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.

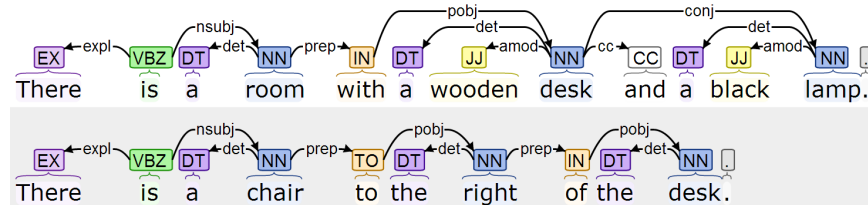


parsing



Scene Generation Pipeline

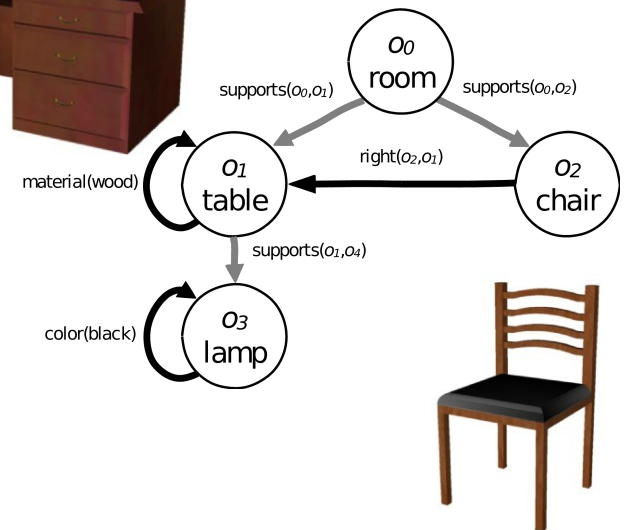
There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.



parsing

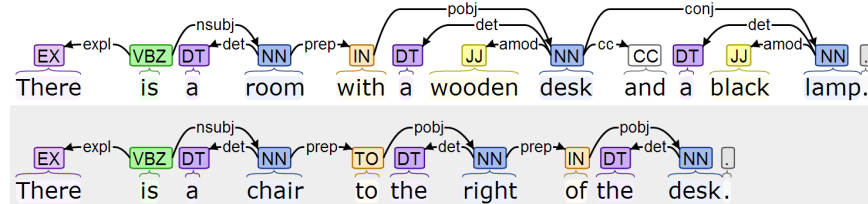


object
selection



Scene Generation Pipeline

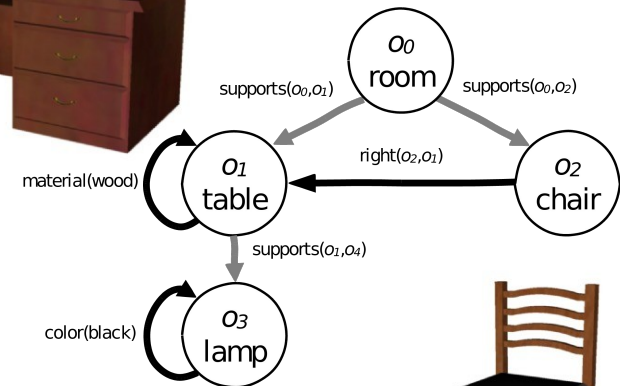
There is a room with a wooden desk and a black lamp. There is a chair to the right of the desk.



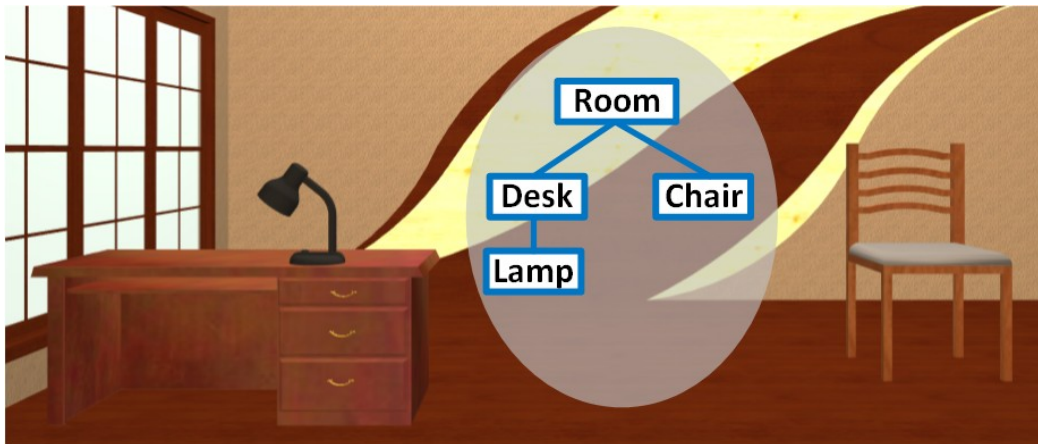
parsing



object
selection

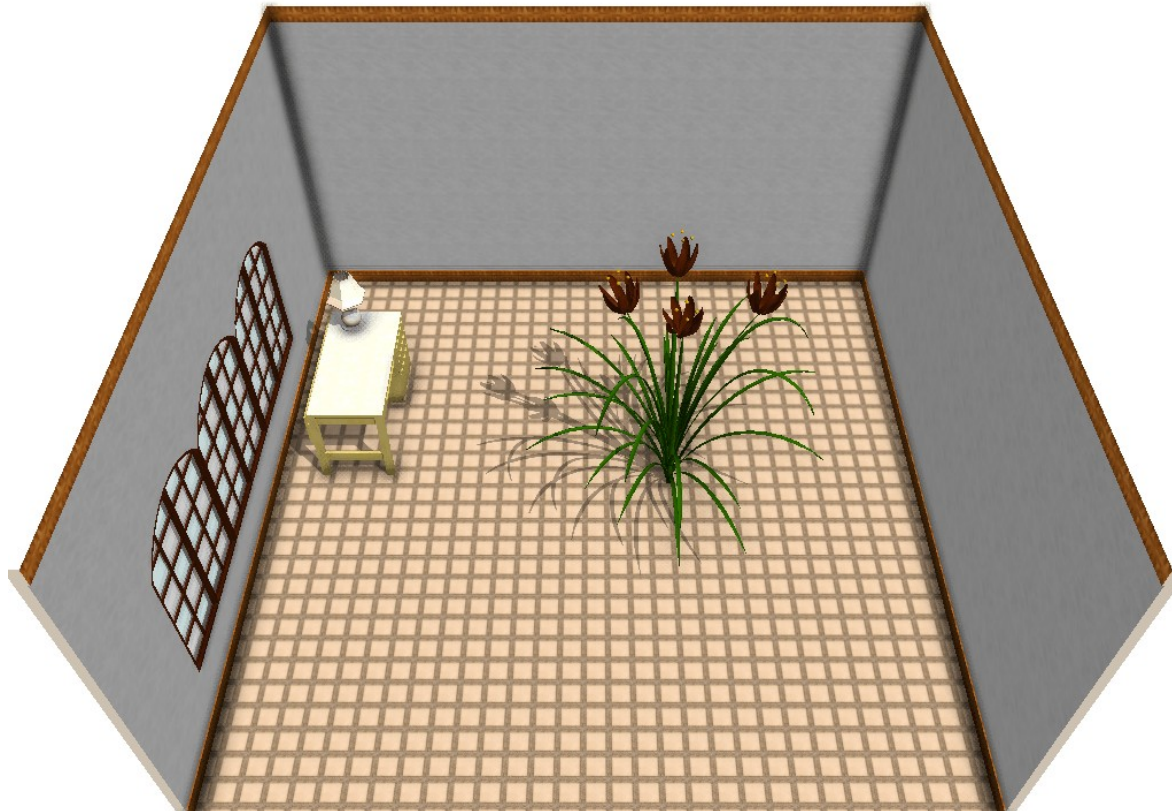


layout



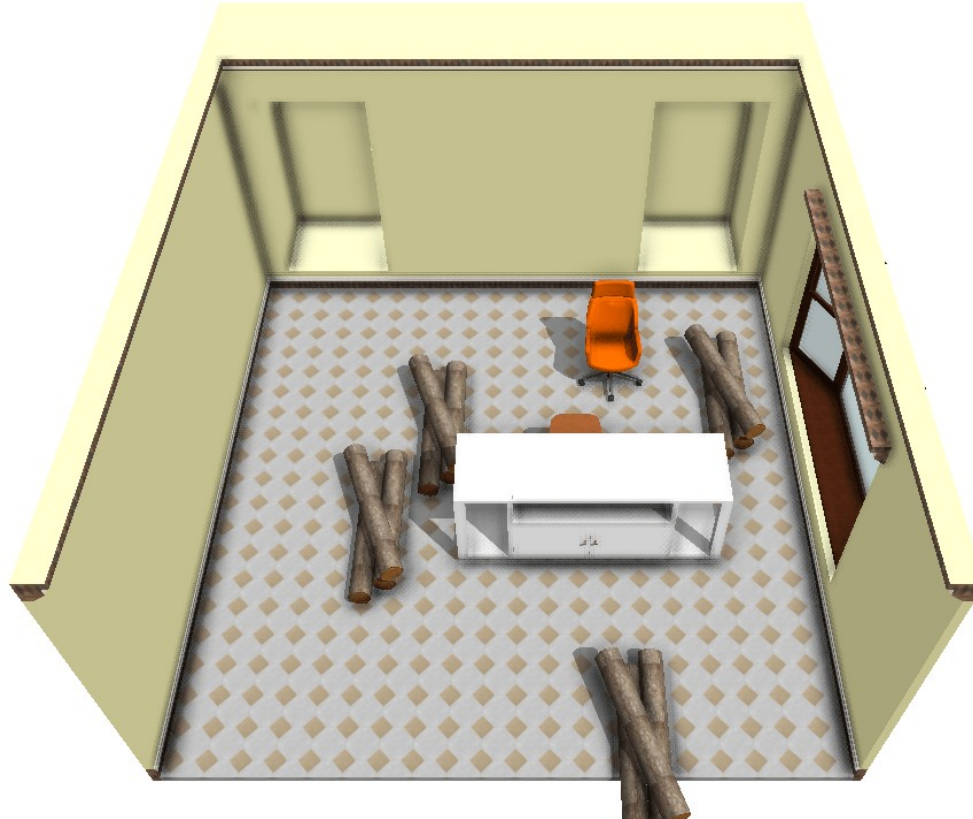
Selected errors

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



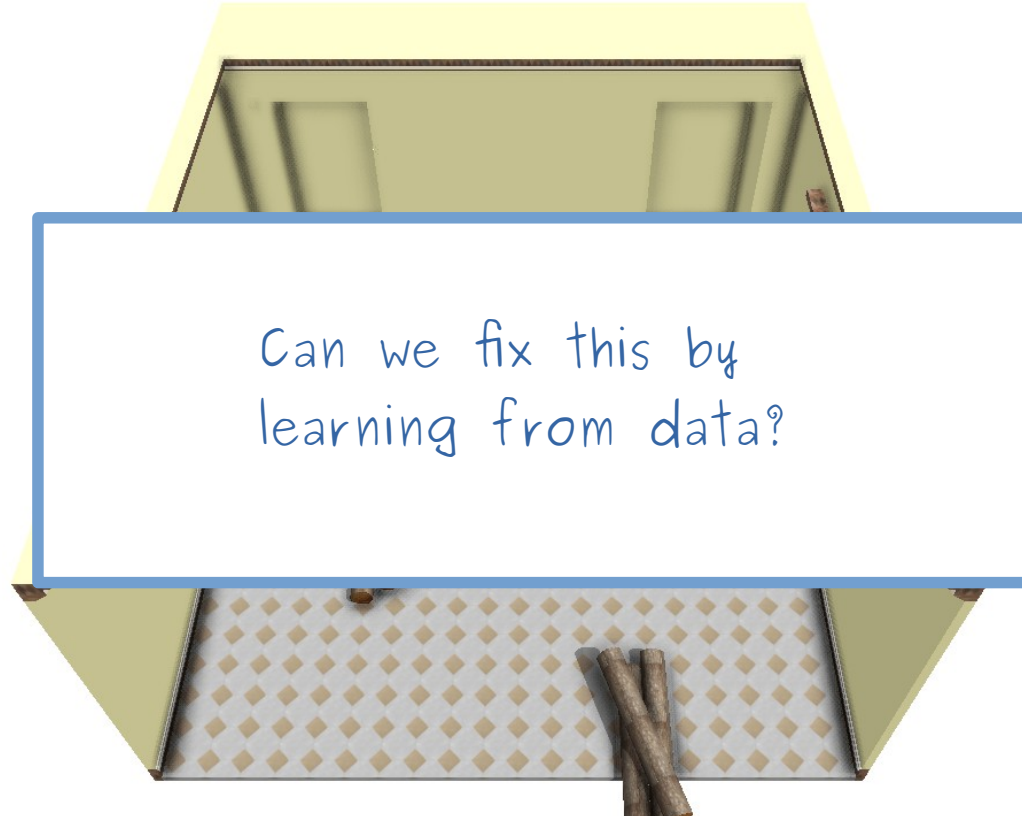
Selected errors

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



Selected errors

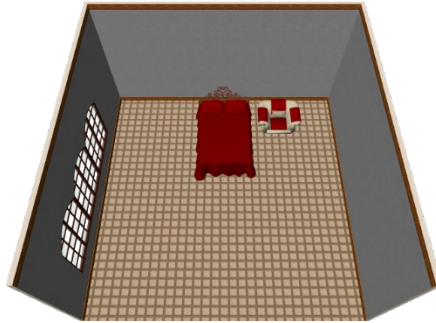
*Wood table and **four wood**
chairs in the center of the room*



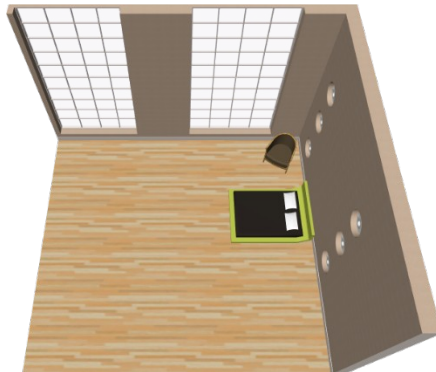
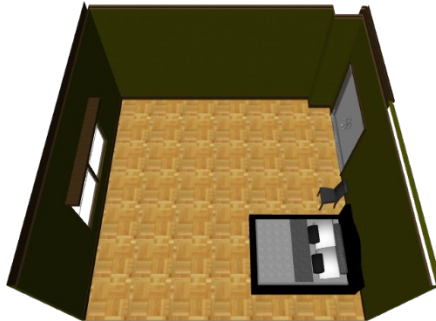
Dataset

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

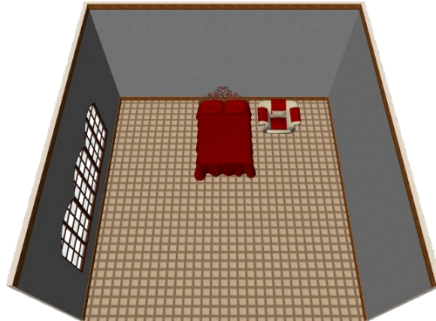
Dataset



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



Dataset

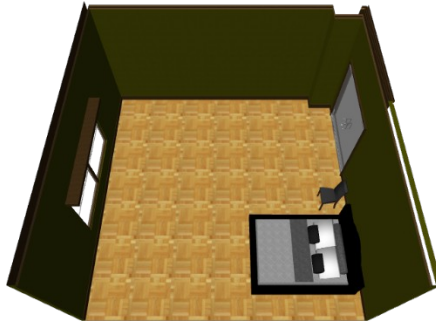


The room has three windows on one wall. There is a red bed in the back of the room. Along side the bed is a side chair that is red and white.

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.

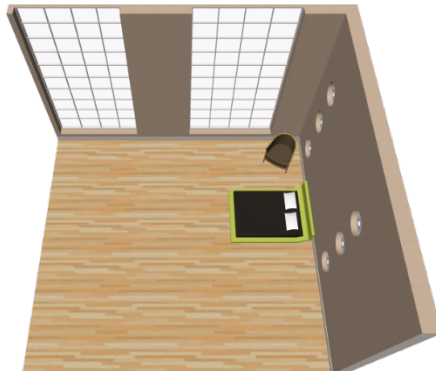
There is a bed and there is a chair next to the bed.



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.

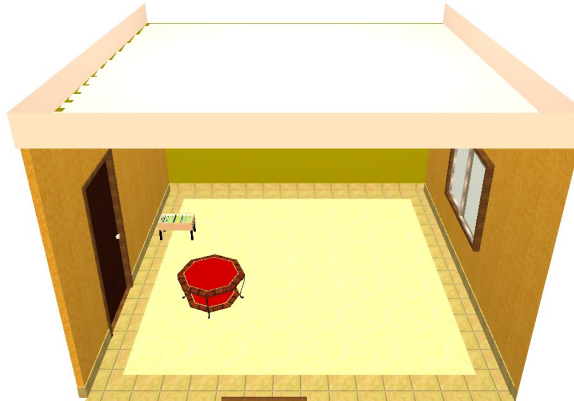
Discrimination task

brown room with a refrigerator in the back corner

A



B



C



D



E



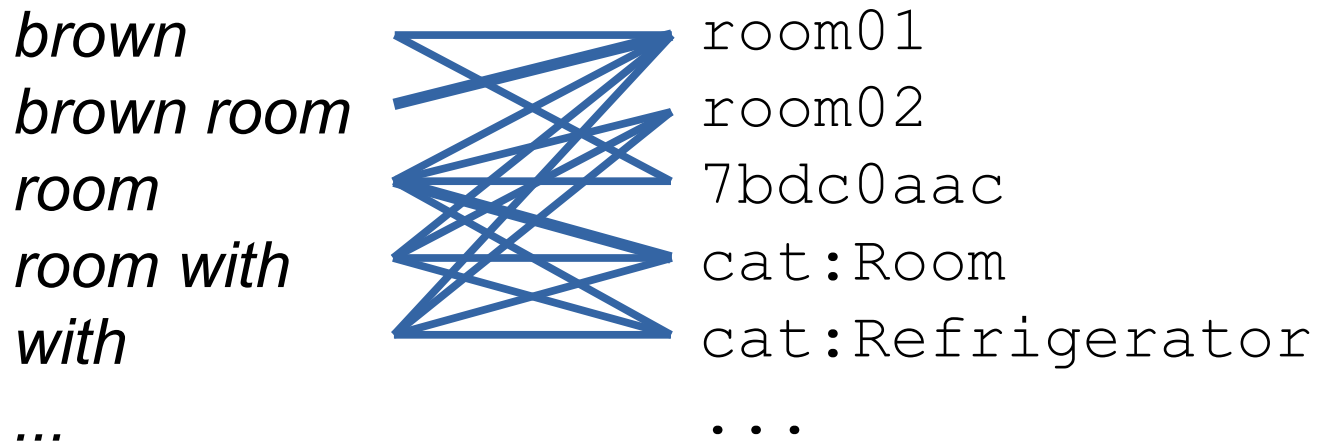
Discrimination task

brown room with a refrigerator in the back corner



Learning lexical items

- One-vs.-all logistic regression
- Features: $\mathbf{1}\{(\text{language}, \text{object})\}$
 - language: bag-of-words / bag-of-bigrams
 - object: model id / category



Discrimination results

- Accuracy (% correct scenes identified)

	Random set	Same seed
Model ids only + unigrams	72.2%	56.7%
Model ids only + bigrams	72.1%	57.4%
Categories only + bigrams	77.4%	46.8%
Both + unigrams	83.5%	63.3%
Both + bigrams	85.0%	64.6%

Generation using learned weights

- Gather all features involving unigrams/bigrams in input
- Group by object, sum weights, choose top k
 - $k = 4$ chosen based on average number of objects in human-constructed scenes
- No relationships enforced between objects

Human evaluation results

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

	Mean score
Random objects	1.68
Learned objects (k=4)	2.61
Rule-based parser	3.15
<i>Human-built</i>	5.87

Human evaluation results

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

	Mean score
Random objects	1.68
Learned objects (k=4)	2.61
Rule-based parser	3.15
<i>Human-built</i>	5.87

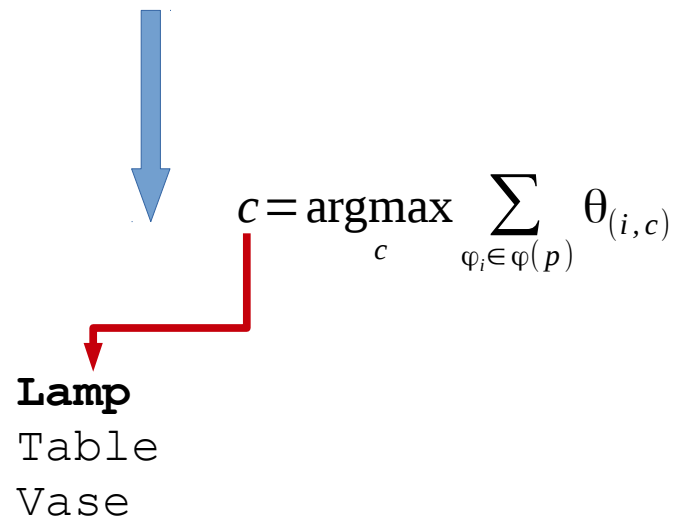


Parsing + learned lexical grounding

there is a room with
a wooden desk and
a **black lamp**

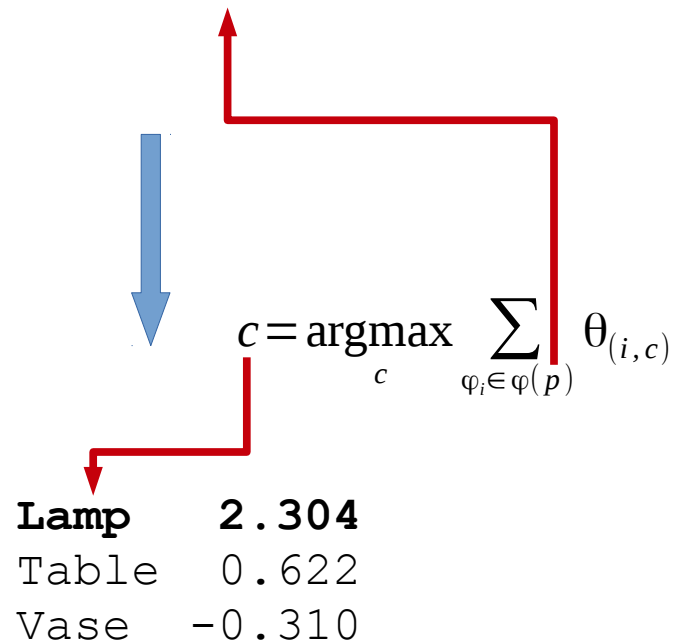
Parsing + learned lexical grounding

there is a room with
a wooden desk and
a **black lamp**



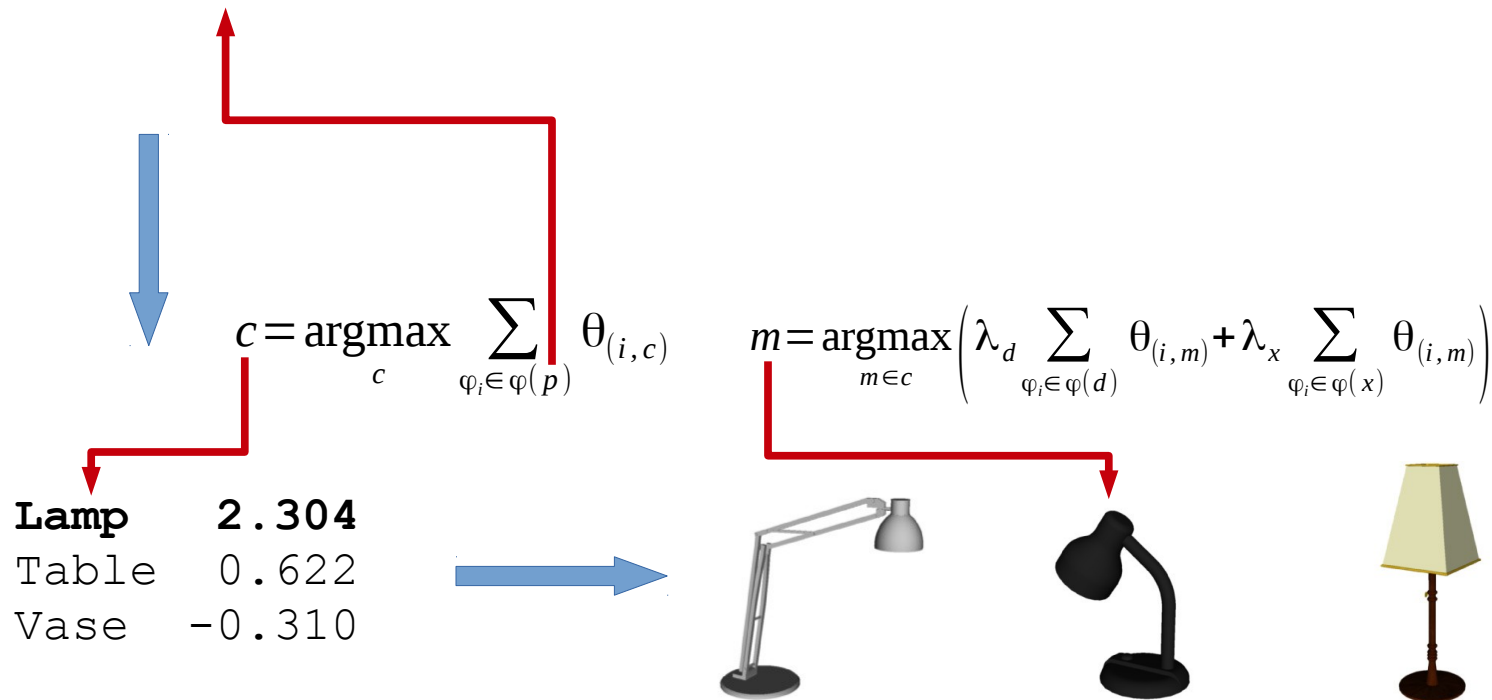
Parsing + learned lexical grounding

there is a room with
a wooden desk and
a **black lamp**



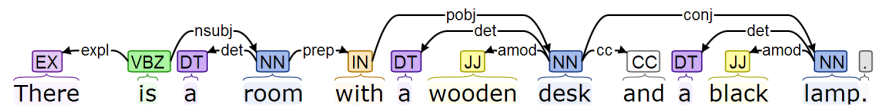
Parsing + learned lexical grounding

there is a room with
a wooden desk and
a **black lamp**



Parsing + learned lexical grounding

there is a room with
a wooden desk and
a **black lamp**



$$c = \operatorname{argmax}_c \sum_{\varphi_i \in \varphi(p)} \theta_{(i,c)}$$

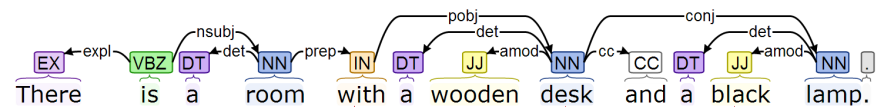
Lamp	2.304
Table	0.622
Vase	-0.310

$$m = \operatorname{argmax}_{m \in c} \left(\lambda_d \sum_{\varphi_i \in \varphi(d)} \theta_{(i,m)} + \lambda_x \sum_{\varphi_i \in \varphi(x)} \theta_{(i,m)} \right)$$



Parsing + learned lexical grounding

there is a room with
a wooden desk and
a **black lamp**



$$c = \operatorname{argmax}_c \sum_{\varphi_i \in \varphi(p)} \theta_{(i,c)}$$

Lamp	2.304
Table	0.622
Vase	-0.310

$$m = \operatorname{argmax}_{m \in c} \left(\lambda_d \sum_{\varphi_i \in \varphi(d)} \theta_{(i,m)} + \lambda_x \sum_{\varphi_i \in \varphi(x)} \theta_{(i,m)} \right)$$



-0.302



0.460



-0.021

Generated scene examples

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.



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.



Human evaluation results

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

	Mean score
Random objects	1.68
Learned objects (k=4)	2.61
Rule-based parser	3.15
Parser + learned	3.73
<i>Human-built</i>	<i>5.87</i>

Human evaluation results

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

	Mean score
Random objects	1.68
Learned objects (k=4)	2.61
Rule-based parser	3.15
Parser + learned	3.73
<i>Human-built</i>	<i>5.87</i>



Missing pieces

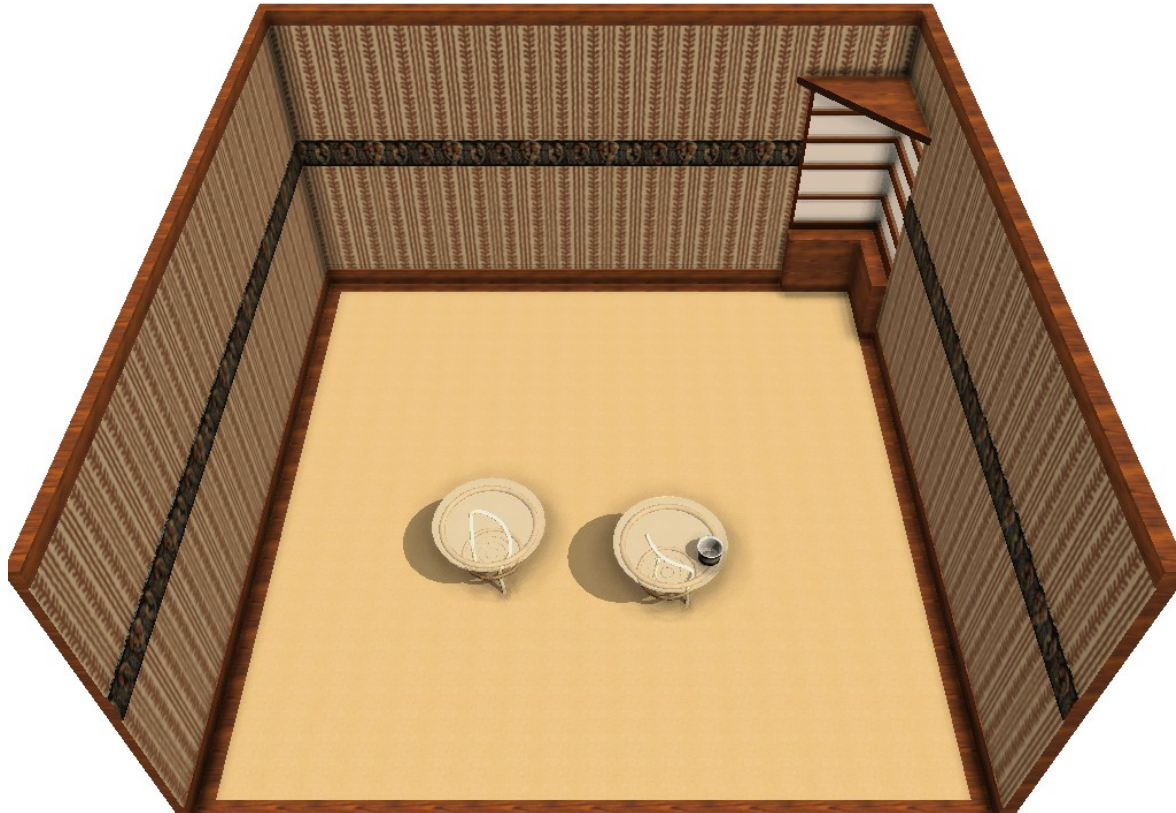
- Compositional meaning
 - Spatial relationships
 - Non-adjacent modifiers
 - Coreference
- Learn geometry and language jointly
- Interaction
- Context-sensitive understanding

Missing pieces

- Compositional meaning
 - Spatial relationships
 - Non-adjacent modifiers
 - **Coreference**
- Learn geometry and language jointly
- Interaction
- Context-sensitive understanding

Coreference

*There in the middle is a **table**.
On the **table** is a cup.*



Missing pieces

- Compositional meaning
 - **Spatial relationships**
 - **Non-adjacent modifiers**
 - Coreference
- Learn geometry and language jointly
- Interaction
- Context-sensitive understanding

A job for semantic parsing?

- Learn semantic representation as a latent variable

```
(EXISTS (and (CAT Refrigerator)
              (BACK_OF room02) ) )
```

- New features this enables
 - structure of semantics
 - type and number of composition rules
 - whole-scene scoring features

Discrimination: scoring scenes

- Constraint satisfaction score:
number of unsatisfied conditions

$$\text{Score}(S; T) = \sum_{p \in T} \max_{s \in S} -\mathbf{1}[\neg p(s)]$$

- “Exhaustivity” score:

$$\text{Exh}(S; T) = \sum_{s \in S} \max_{p \in T} -\mathbf{1}[\neg p(s)]$$

Discrimination: scoring scenes

- Constraint satisfaction score:
number of unsatisfied conditions

$$\text{Score}(S; T) = \sum_{p \in T} \max_{s \in S} -\mathbf{1}[\neg p(s)]$$

- “Exhaustivity” score:

$$\text{Exh}(S; T) = \sum_{s \in S} \max_{p \in T} -\mathbf{1}[\neg p(s)]$$

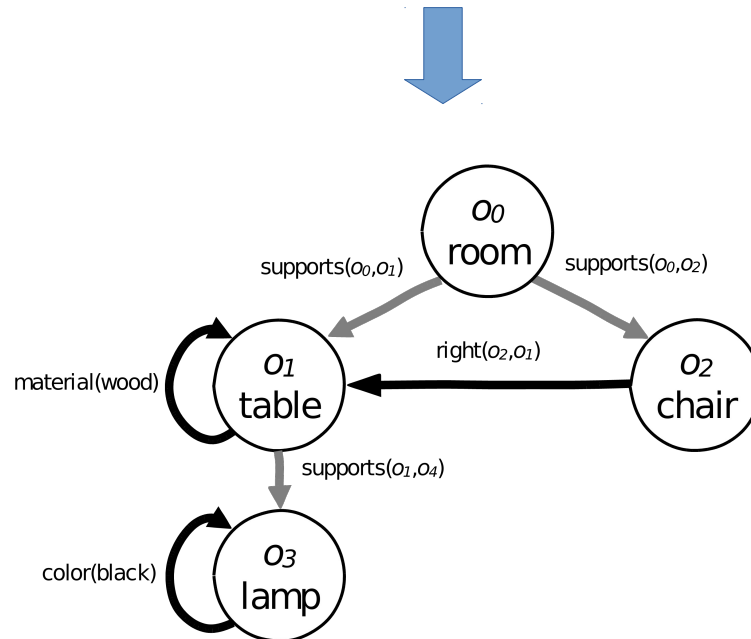
“1. Make your contribution as informative as is required...

“2. Do not make your contribution more informative than is required.”

Paul Grice (1989)

Generation

```
(EXISTS (and (CAT Chair) (RIGHT_OF  
(and (CAT Table) (SUPPORTING  
(CAT Lamp))))))
```



Want to learn more?



Angel Chang
angelx@



Manolis Savva
msavva@



Christopher Manning
manning@



Will Monroe
wmonroe4@





Christopher Potts
cgpotts@

└ Also available at office hours / ┐
right here after class

Appendix: MaxEnt model

- One-vs.-all logistic regression


$$h_{\theta}(y|x) = \frac{1}{1 + \exp[-\varphi(x, y)^T \theta]}$$

$$\hat{y}^{(i)} = \arg \max_{y \in Y^{(i)}} h_{\theta}(y|x^{(i)})$$

$$J(\theta) = \sum_{i=1}^m \sum_{y \in Y^{(i)}} (\mathbf{1}\{y = y^{(i)}\} \log h_{\theta}(y|x^{(i)}) + \mathbf{1}\{y \neq y^{(i)}\} \log [1 - h_{\theta}(y|x^{(i)})])$$

Appendix: Semantic parsing model

- Structured prediction

logical formula

$$h_{\theta}(z|x) = \frac{\exp[-\varphi(x, z)^T \theta]}{\sum_{z' \in D(x; \theta)} \exp[-\varphi(x, z')^T \theta]}$$

beam search

$$\llbracket z \rrbracket = \arg \max_{y \in Y^{(i)}} \text{Score}(y, z)$$

$$\hat{y}^{(i)} = \llbracket \arg \max_{z \in D(x^{(i)}; \theta)} h_{\theta}(z|x^{(i)}) z \rrbracket$$

$$J(\theta) = \sum_{i=1}^m \log \sum_{z \in D(x^{(i)}; \theta)} \mathbf{1}\{\llbracket z \rrbracket = y^{(i)}\} h_{\theta}(y|x^{(i)})$$

Appendix: Bag-of-objects grammar

```
(rule $Object ($LEMMA_PHRASE) (SimpleLexiconFn (type @Any)))
(rule $Relation ($LEMMA_PHRASE)
  (SimpleLexiconFn (type (-> @Any (-> @Any @Any))))))

(rule $Padding ($PHRASE) (IdentityFn))

(rule $NounPhrase ($Object) (IdentityFn))
(rule $Attribute ($Relation ($Padding optional) $Object)
  (JoinFn binary, unary unaryCanBeArg1 betaReduce))
(rule $NounPhrase ($Object ($Padding optional) $Attribute)
  (JoinFn unary, binary unaryCanBeArg1 betaReduce))

(rule $Fact ($NounPhrase) (JoinFn (arg0 EXISTS) unaryCanBeArg1))
(rule $Facts ($Fact) (IdentityFn))
(rule $Facts ($Fact ($Padding optional) $Facts) (MergeFn and))

(rule $ROOT (($Padding optional) $Facts ($Padding optional))
  (IdentityFn))
```