Understanding vision through language and language through vision

CENTER FOR
Brains
Minds+
Machines

Andrei Barbu

Language and perception

Language and perception

Caption, answer questions, understand a description, explain it to someone, engage in a conversation, give agents commands, imagine something different, recognize its description in a story, rewrite that description in another language, understand if someone is missing the point, reproduce it, intervene, etc.

Recognition

Recognition

Retrieval

Recognition

Retrieval

Generation

Recognition

Retrieval

Generation

Question answering

Recognition

Retrieval

Generation

Question answering

Disambiguation

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Follow commands

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Follow commands

Paraphrasing

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Follow commands

Paraphrasing

Translation

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Follow commands

Paraphrasing

Translation

Common sense reasoning

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Follow commands

Paraphrasing

Translation

Common sense reasoning

Planning

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Follow commands

Paraphrasing

Translation

Common sense reasoning

Planning

…

Recognition

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Follow commands

Paraphrasing

Translation

Common sense reasoning

Planning

...

Computer vision

NLP

Robotics

AI

Recognition

$P(\text{sentence}, \text{video})$

Narayanaswamy *et al.* 2014

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Follow commands

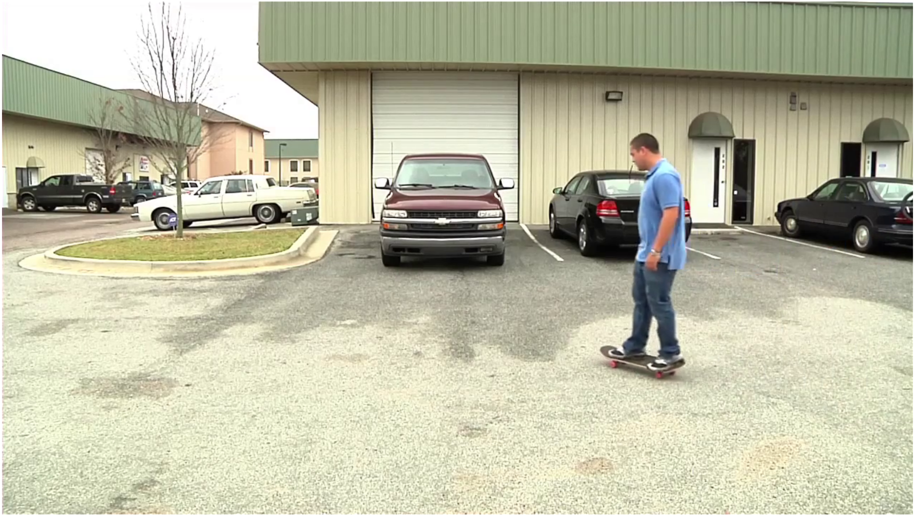
Paraphrasing

Translation

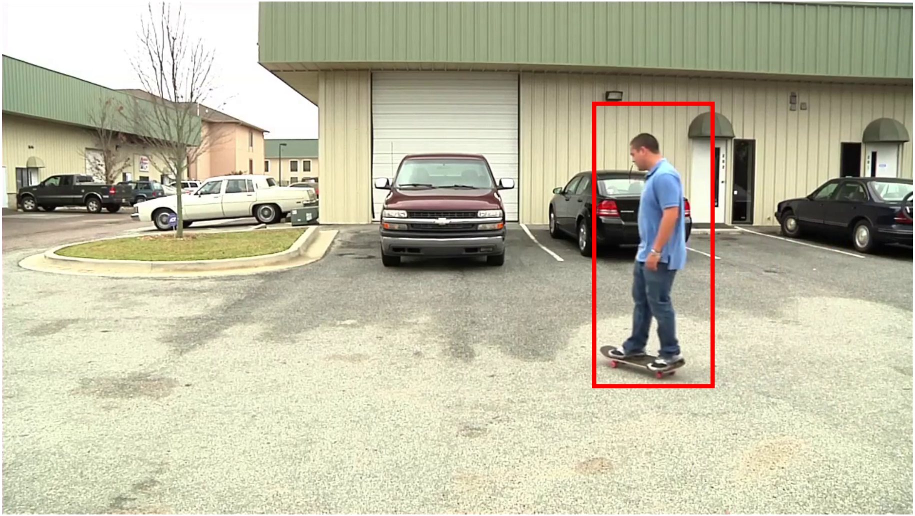
Common sense reasoning

Planning

…

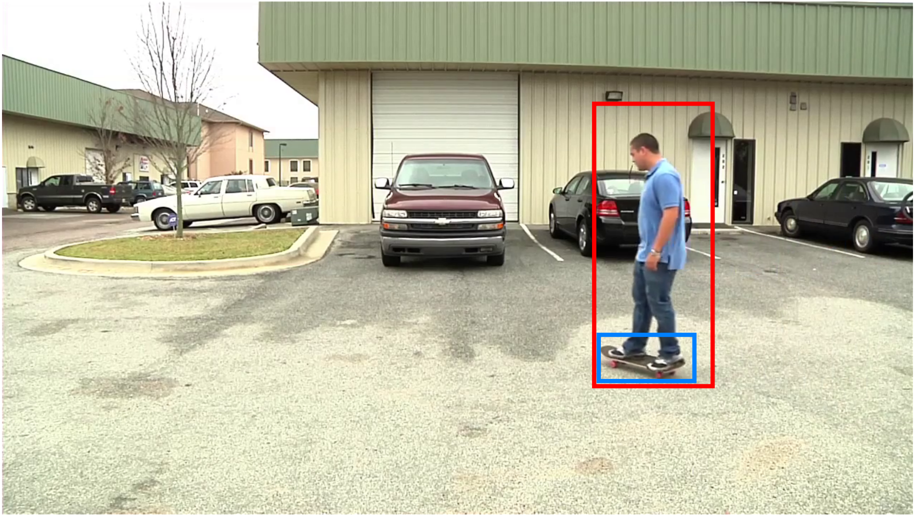


The person rode the skateboard leftward.



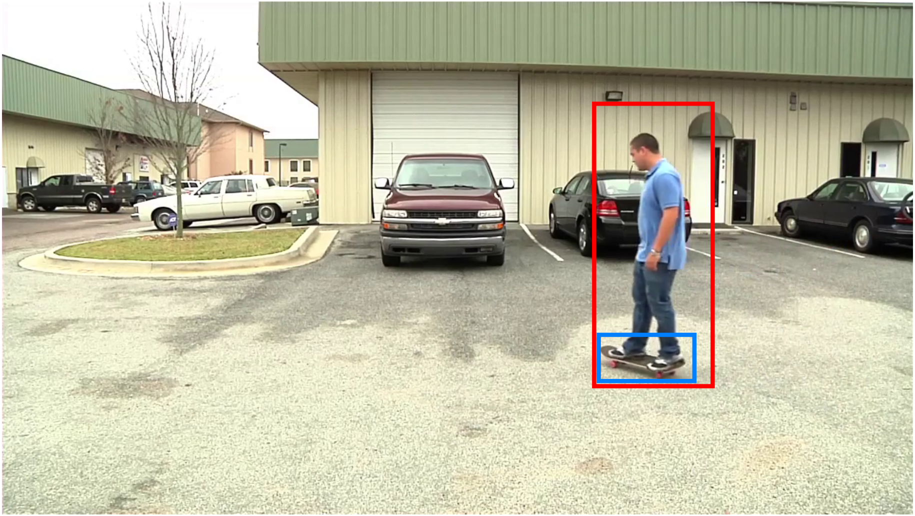
The person rode the skateboard leftward.

object detector, tracker, event recognizer



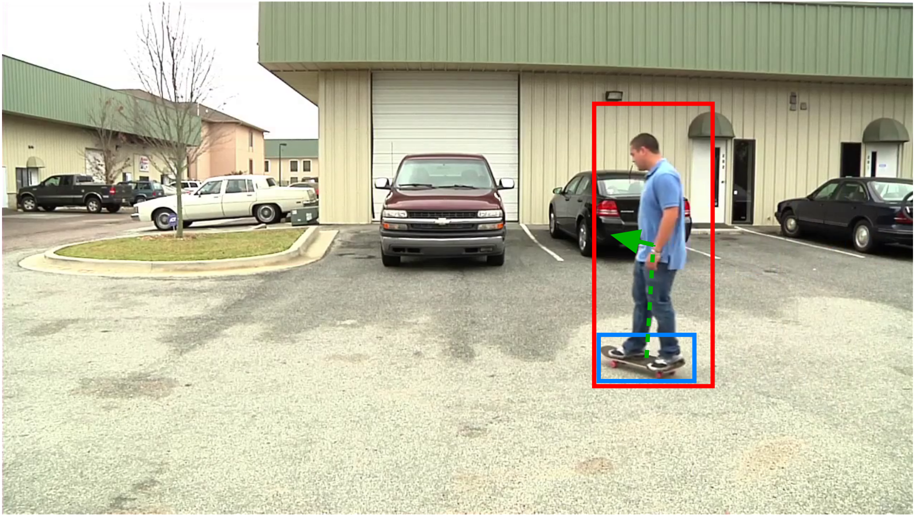
The person rode the skateboard leftward.

object detector, tracker, event recognizer



The person rode the skateboard leftward.

object detector, tracker, event recognizer



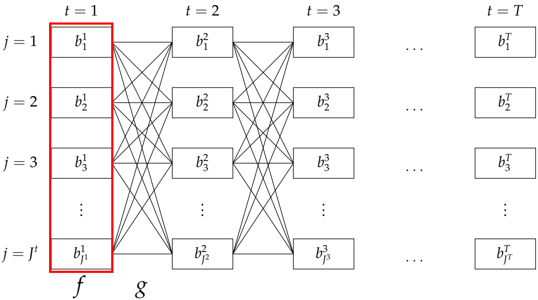
The person rode the skateboard leftward.

object detector, tracker, event recognizer

Tracking with higher-level knowledge

Barbu *et al.* 2012

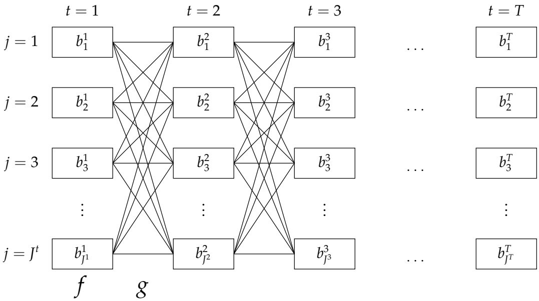
Tracking with higher-level knowledge



detection / object / frame

Barbu *et al.* 2012

Tracking with higher-level knowledge

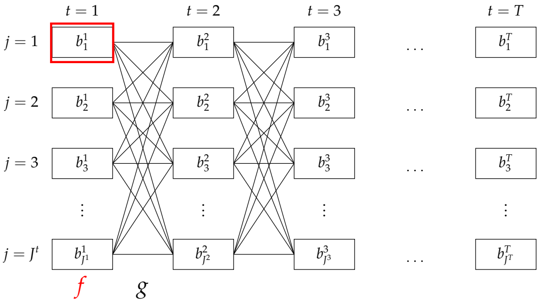


detection / object / frame

temporally coherent track

Barbu *et al.* 2012

Tracking with higher-level knowledge



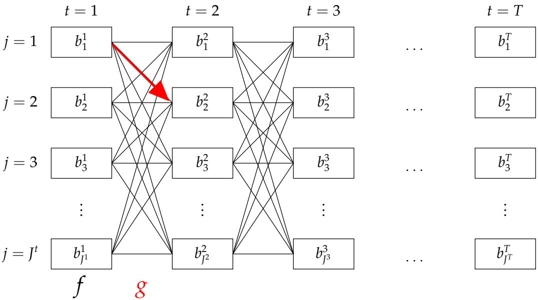
detection / object / frame

temporally coherent track

- object detector confidence (f)

Barbu *et al.* 2012

Tracking with higher-level knowledge



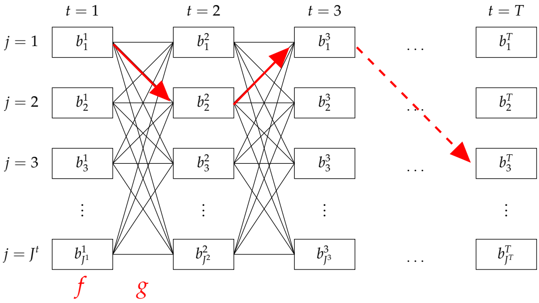
detection / object / frame

temporally coherent track

- object detector confidence (f)
- motion coherence (g)

Barbu *et al.* 2012

Tracking with higher-level knowledge



detection / object / frame

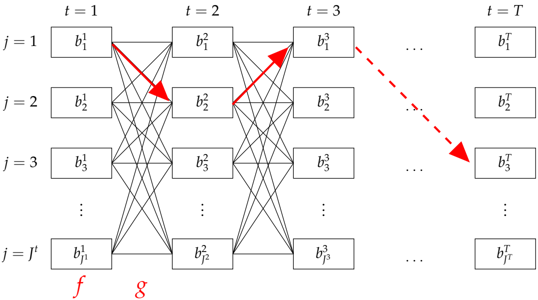
temporally coherent track

- object detector confidence (f)
- motion coherence (g)

optimal path through the lattice of detections

Barbu *et al.* 2012

Tracking with higher-level knowledge



detection / object / frame

temporally coherent track

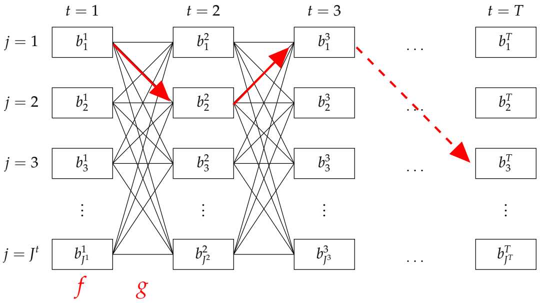
- object detector confidence (f)
- motion coherence (g)

optimal path through the lattice of detections

$$\max_{j^1,\ldots,j^T} \sum_{t=1}^{T} f(b_{j^t}^t) + \sum_{t=2}^{T} g(b_{j^{t-1}}^{t-1}, b_{j^t}^t)$$

Barbu *et al.* 2012

Tracking with higher-level knowledge



detection / object / frame

temporally coherent track

- object detector confidence (f)
- motion coherence (g)

optimal path through the lattice of detections

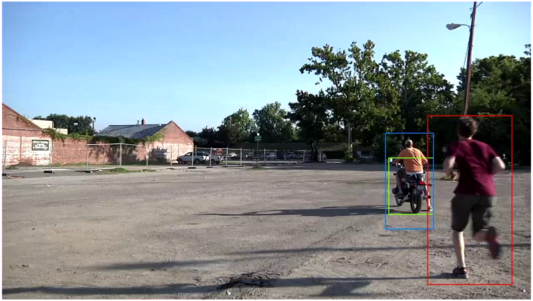
dynamic programming

Bellman (1957), Viterbi (1967)

$$\max_{j^1,\ldots,j^T} \sum_{t=1}^{T} f(b^t_{j^t}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}}, b^t_{j^t})$$

Barbu *et al.* 2012

Tracking with higher-level knowledge



detection / object / frame

temporally coherent track

- object detector confidence (f)
- motion coherence (g)

optimal path through the lattice of detections

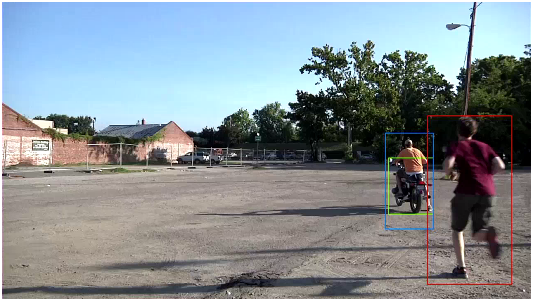
dynamic programming

Bellman (1957), Viterbi (1967)

$$\max_{j^1,\ldots,j^T} \sum_{t=1}^{T} f(b^t_{j^t}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}}, b^t_{j^t})$$

Barbu *et al.* 2012

Tracking with higher-level knowledge



detection / object / frame

temporally coherent track

- object detector confidence (f)
- motion coherence (g)

optimal path through the lattice of detections

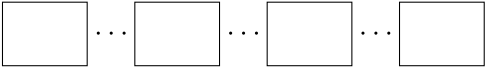
dynamic programming

Bellman (1957), Viterbi (1967)

$$\max_{j^1,\ldots,j^T} \sum_{t=1}^{T} f(b^t_{j^t}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}}, b^t_{j^t})$$

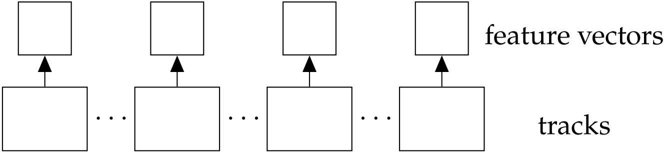
Barbu *et al.* 2012

Event recognition

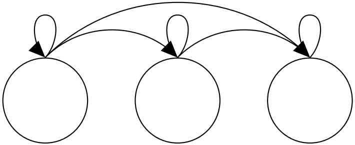


tracks

Event recognition

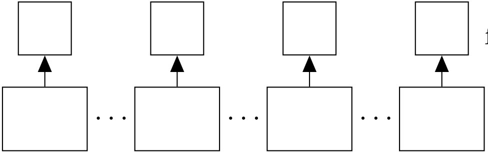


Event recognition



HMMs

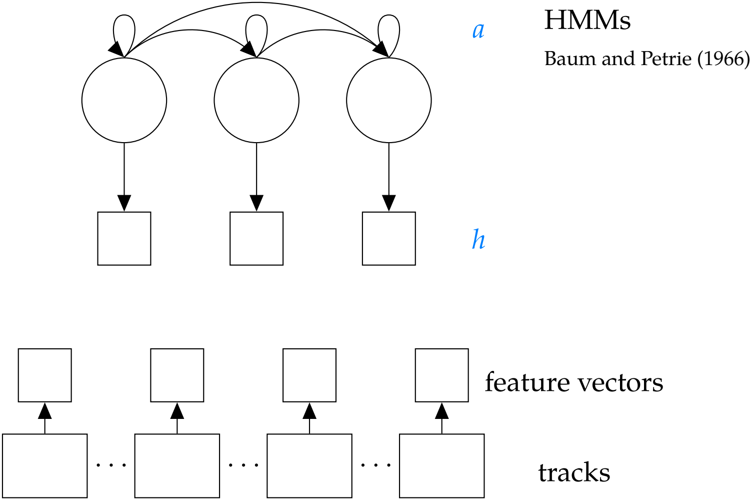
Baum and Petrie (1966)



feature vectors

tracks

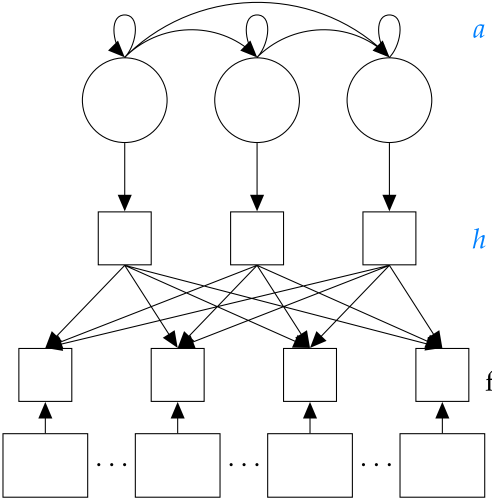
Event recognition



HMMs

Baum and Petrie (1966)

Event recognition



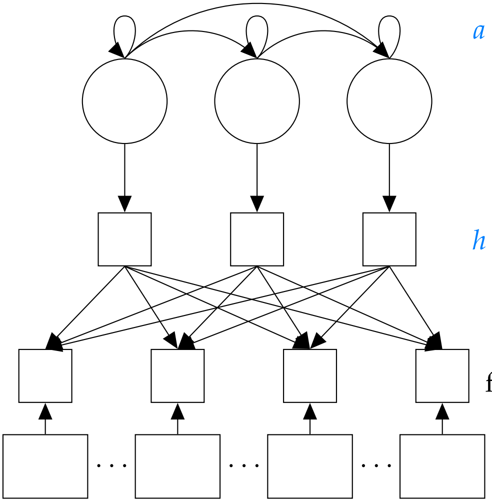
HMMs

Baum and Petrie (1966)

feature vectors

tracks

Event recognition



feature vectors

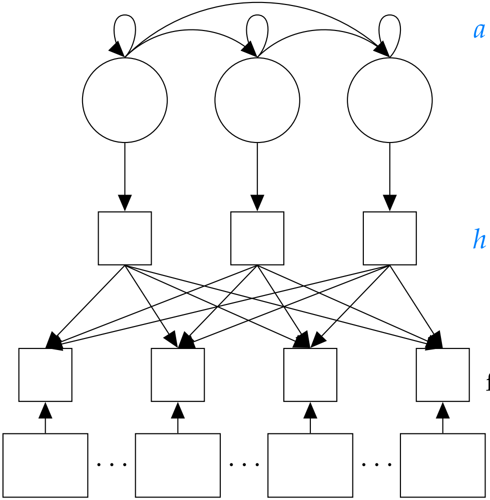
tracks

HMMs

Baum and Petrie (1966)

One HMM per event class

Event recognition



a

h

feature vectors

tracks

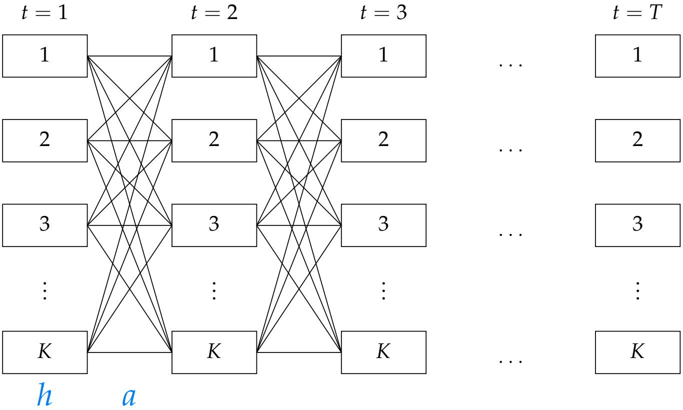
HMMs

Baum and Petrie (1966)

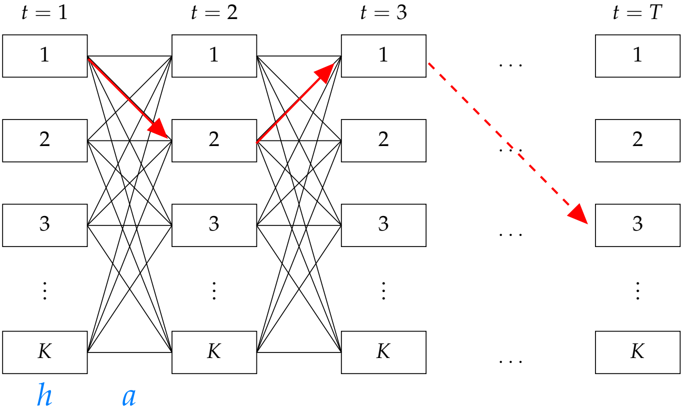
One HMM per event class

Try each HMM

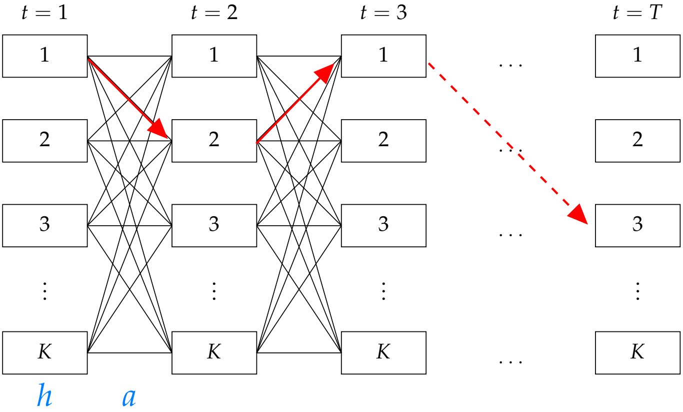
Event recognition



Event recognition



Event recognition



$$\max_{k^1,\ldots,\, k^T} \sum_{t=1}^{T} h(k^t, b^t_{\hat{\jmath}^t}) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

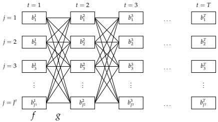
Building sentences out of trackers and words

Siddharth *et al.* 2014

Building sentences out of trackers and words

Viterbi tracker

track 1



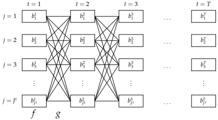
$$\max_{j_1^1,\ldots,\, j_1^T} \qquad \sum_{t=1}^{T} f(b^t_{j_l^t}) + \sum_{t=2}^{T} g(b^{t-1}_{j_l^{t-1}}, b^t_{j_l^t})$$

Siddharth *et al.* 2014

Building sentences out of trackers and words

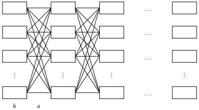
Event tracker for intransitive verbs

track 1



t = 1
t = 2
t = 3
t = T
j = 1
j = 2
j = 3
j = J^t
b_1^1 b_1^2 b_1^3 b_1^T
b_2^1 b_2^2 b_2^3 b_2^T
b_3^1 b_3^2 b_3^3 b_3^T
$b_{J^1}^1$ $b_{J^2}^2$ $b_{J^3}^3$ $b_{J^T}^T$
f
g

×



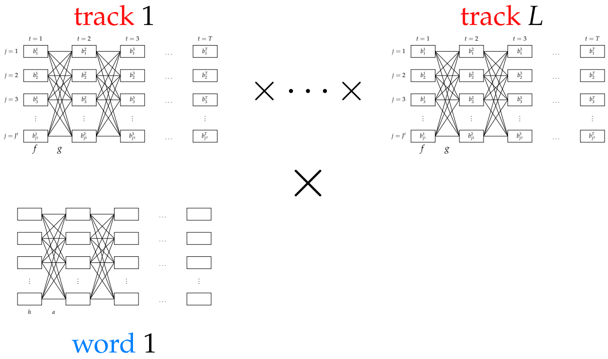
word 1

$$\max_{j_1^1,\ldots,\, j_1^T} \max_{k_1^1,\ldots,\, k_1^T} \sum_{t=1}^{T} f(b^t_{j^t_l}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}_l}, b^t_{j^t_l}) + \sum_{t=1}^{T} h(k^t, b^t_{\hat{j}^t}) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Siddharth *et al.* 2014

Building sentences out of trackers and words

Event tracker



track 1 × ··· × track L

×

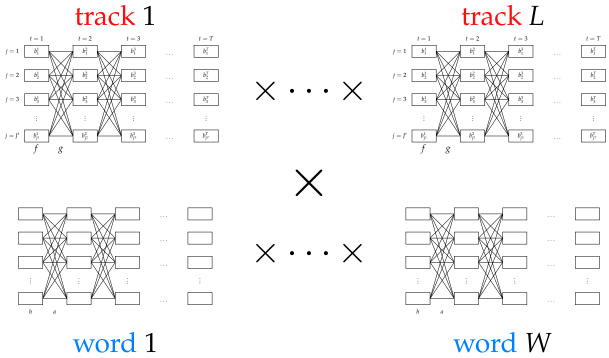
word 1

$$\max_{\substack{j_1^1,\ldots,\, j_1^T\\ \vdots \\ j_L^1,\ldots,\, j_L^T}} \max_{k_1^1,\ldots,\, k_1^T} \sum_{l=1}^{L}\sum_{t=1}^{T} f(b^t_{j^t_l}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}_l}, b^t_{j^t_l}) + \sum_{t=1}^{T} h(k^t, b^t_{\hat{j}^t_{\theta 1}}, b^t_{\hat{j}^t_{\theta 2}}) + \sum_{t=2}^{T} a(k^{t-1}, k^t)$$

Siddharth *et al.* 2014

Building sentences out of trackers and words

Sentence tracker



track 1

track L

word 1

word W

$$\max_{\substack{j_1^1,\ldots,\, j_1^T\\ \vdots \\ j_L^1,\ldots,\, j_L^T}} \; \max_{\substack{k_1^1,\ldots,\, k_1^T\\ \vdots \\ k_W^1,\ldots,\, k_W^T}} \sum_{l=1}^{L}\sum_{t=1}^{T} f(b^t_{j^t_l}) + \sum_{t=2}^{T} g(b^{t-1}_{j^{t-1}_l}, b^t_{j^t_l}) + \sum_{w=1}^{W}\sum_{t=1}^{T} h_w(k^t_w, b^t_{j^t_{\theta^1_w}}, b^t_{j^t_{\theta^2_w}}) + \sum_{t=2}^{T} a_w(k^{t-1}_w, k^t_w)$$

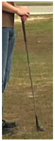
Siddharth *et al.* 2014

Sentence recognizers

The tall person quickly rode the horse leftward away from the other horse.

Sentence recognizers

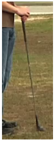
The tall person quickly rode the horse leftward away from the other horse.



…

Sentence recognizers

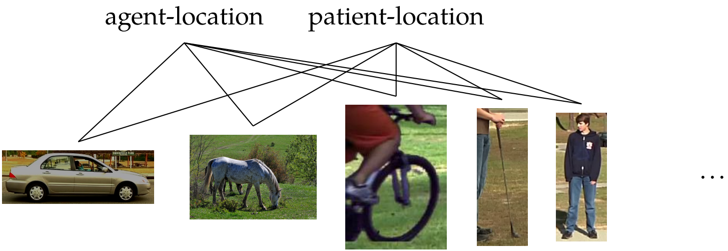
The tall person quickly rode the horse leftward away from the other horse.



. . .

Sentence recognizers

The tall person quickly rode the horse leftward away from the other horse.

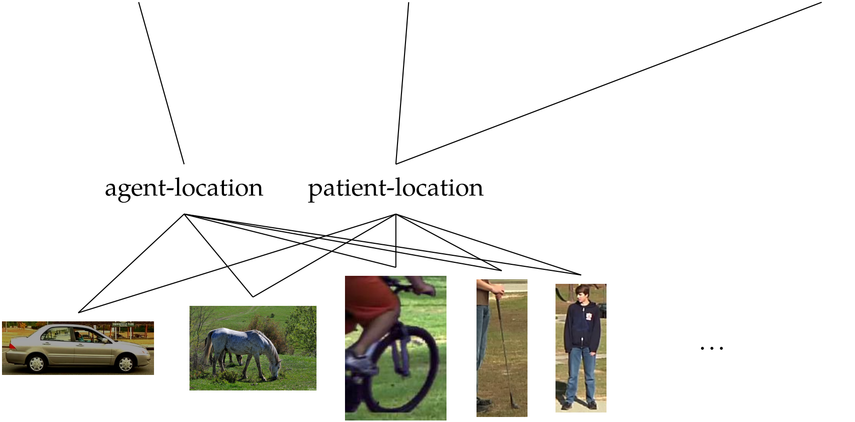


agent-location patient-location

. . .

Sentence recognizers

The tall person quickly rode the horse leftward away from the other horse.

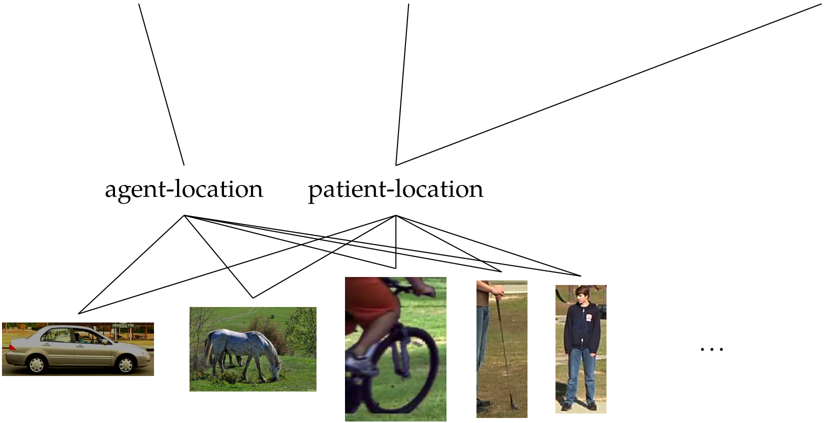


agent-location patient-location

…

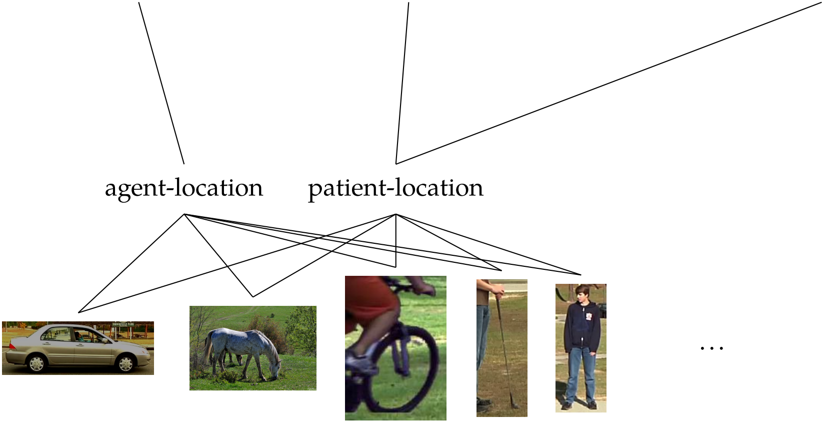
Sentence recognizers

The tall person quickly rode the horse leftward away from the other horse.



Sentence recognizers

The tall person quickly rode the horse leftward away from the other horse.

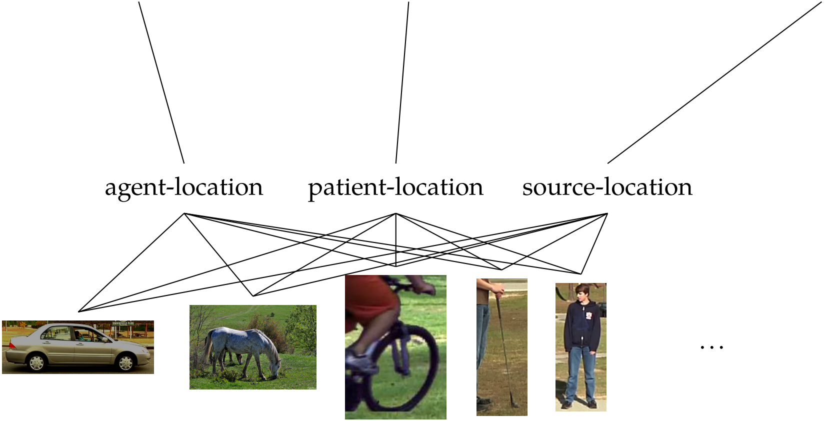


agent-location patient-location

…

Sentence recognizers

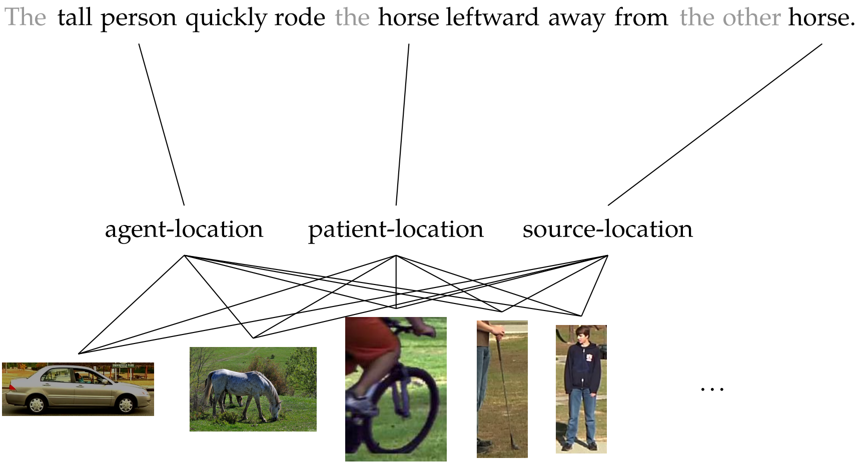
The tall person quickly rode the horse leftward away from the other horse.



agent-location patient-location source-location

…

Sentence recognizers

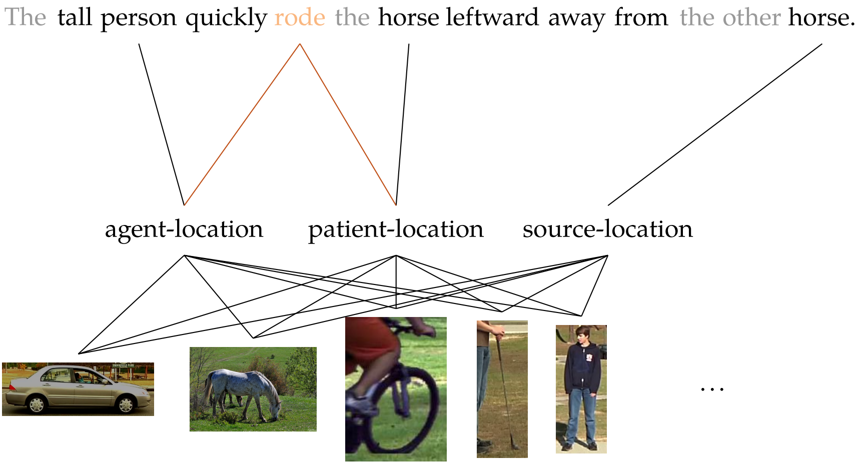


The tall person quickly rode the horse leftward away from the other horse.

agent-location patient-location source-location

…

Sentence recognizers

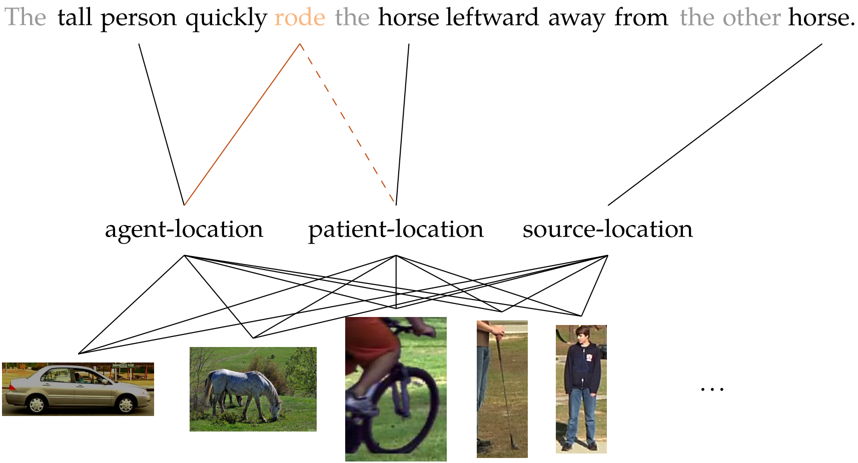


The tall person quickly rode the horse leftward away from the other horse.

agent-location patient-location source-location

. . .

Sentence recognizers

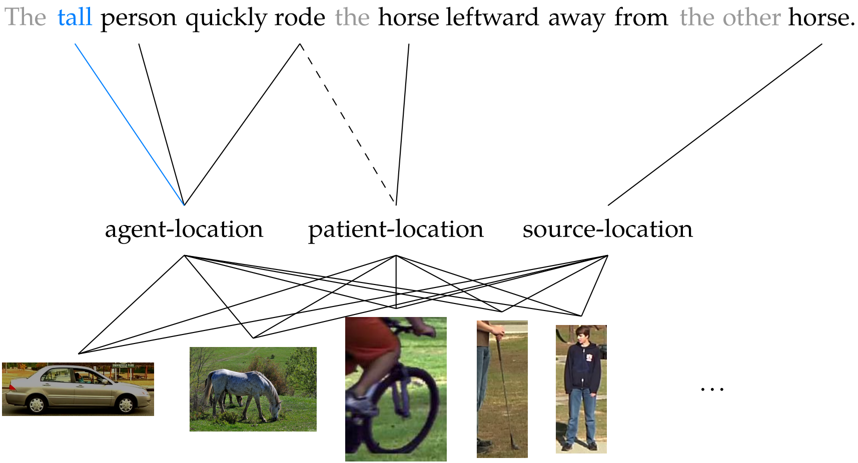


The tall person quickly rode the horse leftward away from the other horse.

agent-location patient-location source-location

…

Sentence recognizers

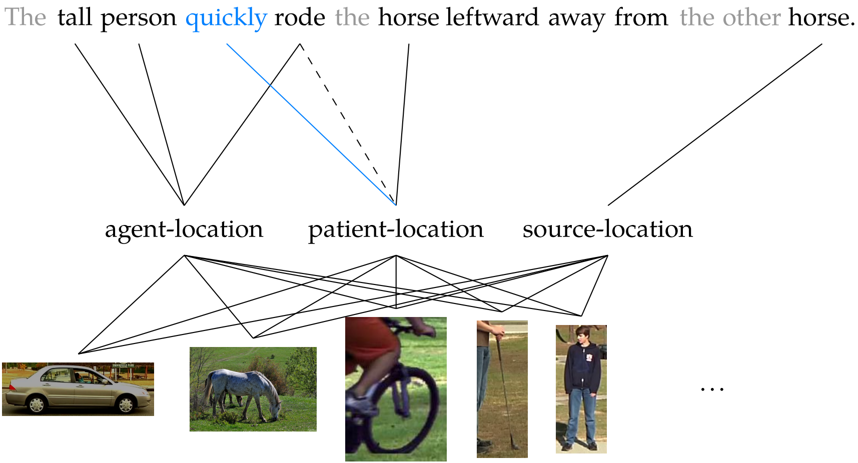


The tall person quickly rode the horse leftward away from the other horse.

agent-location patient-location source-location

…

Sentence recognizers

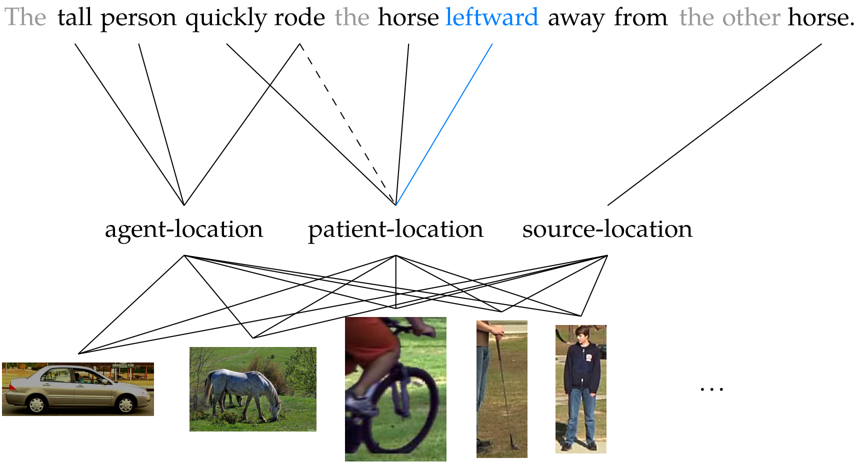


The tall person quickly rode the horse leftward away from the other horse.

agent-location patient-location source-location

…

Sentence recognizers



The tall person quickly rode the horse leftward away from the other horse.

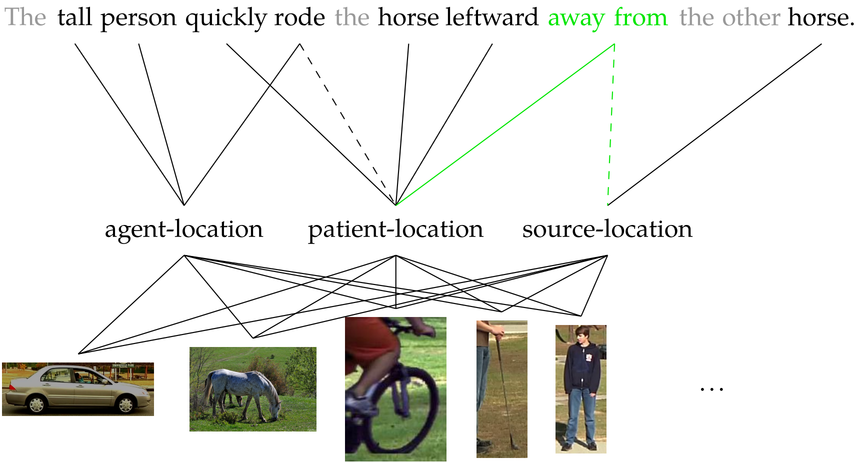
agent-location

patient-location

source-location

…

Sentence recognizers



The tall person quickly rode the horse leftward away from the other horse.

agent-location patient-location source-location

…

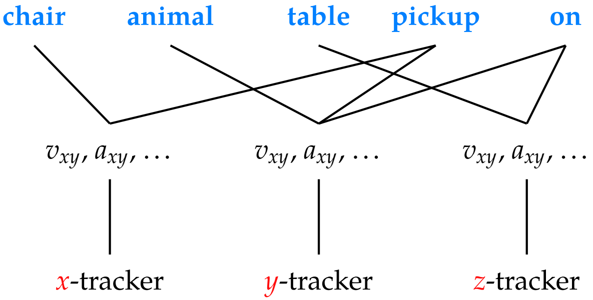
The person picked up the animal from the table.

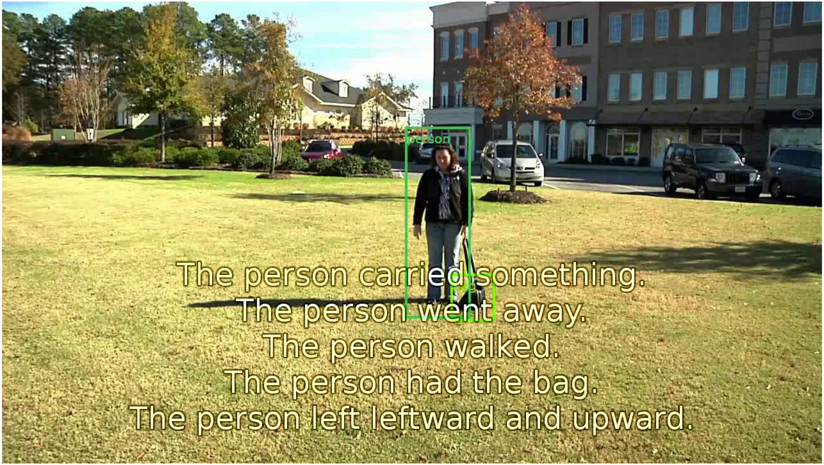
The person picked up the animal from the table.

$$\exists xyz\ \mathbf{person}(x), \mathbf{animal}(y), \mathbf{table}(z), \mathbf{pickup}(x, y), \mathbf{on}(y, z)$$

The person picked up the animal from the table.

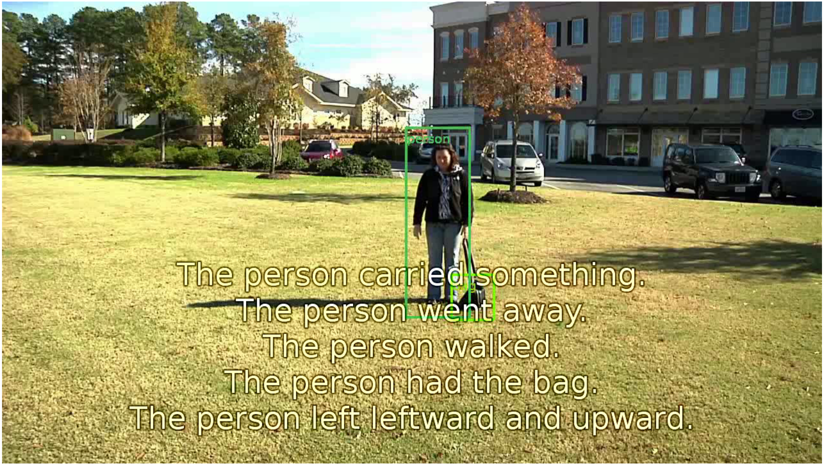
$\exists xyz$ **person**(x), **animal**(y), **table**(z), **pickup**(x, y), **on**(y, z)





The person carried something.
The person went away.
The person walked.
The person had the bag.
The person left leftward and upward.

Barbu *et al.* 2012



The person carried something.
The person went away.
The person walked.
The person had the bag.
The person left leftward and upward.

Barbu *et al.* 2012

Recognition $P(\text{sentence}, \text{video})$ Narayanaswamy *et al.* 2014

Retrieval

Generation

Question answering

Disambiguation

Language acquisition

Follow commands

Paraphrasing

Translation

Common sense reasoning

Planning

…

Recognition $P(\text{sentence}, \text{video})$ Narayanaswamy *et al.* 2014

Retrieval $\underset{v \in V}{\operatorname{argmax}}\ P(s, v)$ Barret *et al.* 2016

Generation

Question answering

Disambiguation

Language acquisition

Follow commands

Paraphrasing

Translation

Common sense reasoning

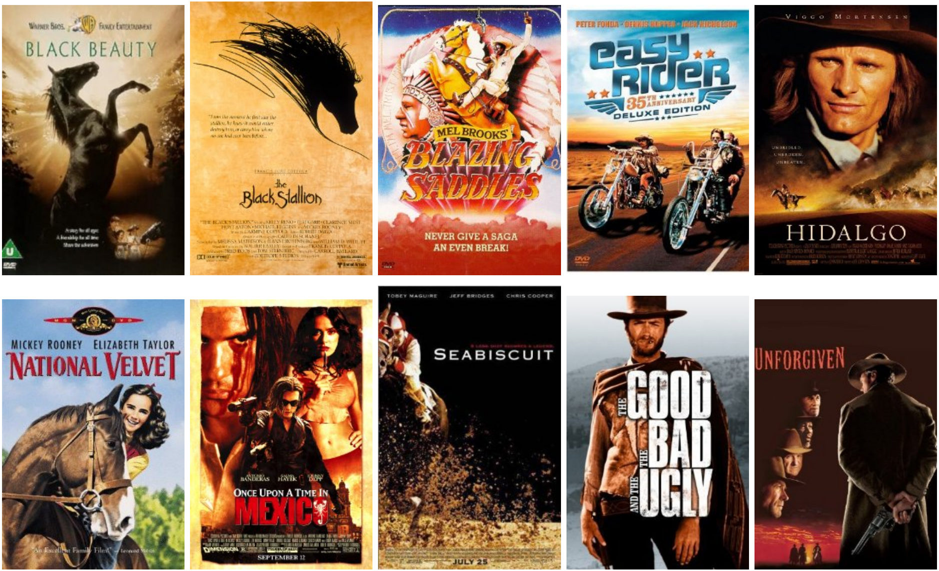
Planning

…

Sentential retrieval

Barret *et al.* 2016

Sentential retrieval



Warner Bros. Family Entertainment
BLACK BEAUTY
the Black Stallion
MEL BROOKS'
BLAZING SADDLES
NEVER GIVE A SAGA AN EVEN BREAK!
PETER FONDA · DENNIS HOPPER · JACK NICHOLSON
easy Rider
35TH ANNIVERSARY
DELUXE EDITION
VIGGO MORTENSEN
HIDALGO
MICKEY ROONEY ELIZABETH TAYLOR
NATIONAL VELVET
ONCE UPON A TIME IN MEXICO
SEPTEMBER 12
TOBEY MAGUIRE JEFF BRIDGES CHRIS COOPER
SEABISCUIT
JULY 25
THE GOOD THE BAD AND THE UGLY
UNFORGIVEN

Barret *et al.* 2016

Recognition $P(\text{sentence}, \text{video})$ Narayanaswamy *et al.* 2014

Retrieval $\operatorname*{argmax}_{v \in V} P(s, v)$ Barret *et al.* 2016

Generation

Question answering

Disambiguation

Language acquisition

Follow commands

Paraphrasing

Translation

Common sense reasoning

Planning

…

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering		
Disambiguation		
Language acquisition		
Follow commands		
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
…		

Generating sentences

Generating sentences

$$
\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{NP\ VP} \\
\mathrm{NP} &\rightarrow \mathrm{D\ [A]\ N\ [PP]} \\
\mathrm{D} &\rightarrow \textit{an} \mid \textit{the} \\
\mathrm{A} &\rightarrow \textit{blue} \mid \textit{red} \\
\mathrm{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object} \\
\mathrm{PP} &\rightarrow \mathrm{P\ NP} \\
\mathrm{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\mathrm{VP} &\rightarrow \mathrm{V\ NP\ [Adv]\ [PP_M]} \\
\mathrm{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down} \\
\mathrm{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\mathrm{PP_M} &\rightarrow \mathrm{P_M\ NP} \\
\mathrm{P_M} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

Generating sentences

$$\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{NP\ VP} \\
\mathrm{NP} &\rightarrow \mathrm{D\ [A]\ N\ [PP]} \\
\mathrm{D} &\rightarrow \textit{an} \mid \textit{the} \\
\mathrm{A} &\rightarrow \textit{blue} \mid \textit{red} \\
\mathrm{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object} \\
\mathrm{PP} &\rightarrow \mathrm{P\ NP} \\
\mathrm{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\mathrm{VP} &\rightarrow \mathrm{V\ NP\ [Adv]\ [PP_M]} \\
\mathrm{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down} \\
\mathrm{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\mathrm{PP_M} &\rightarrow \mathrm{P_M\ NP} \\
\mathrm{P_M} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}$$

147,123,874,800 sentences without recursion

Generating sentences

$$
\begin{aligned}
\text{S} &\rightarrow \text{NP VP} \\
\text{NP} &\rightarrow \text{D [A] N [PP]} \\
\text{D} &\rightarrow \textit{an} \mid \textit{the} \\
\text{A} &\rightarrow \textit{blue} \mid \textit{red} \\
\text{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object} \\
\text{PP} &\rightarrow \text{P NP} \\
\text{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\text{VP} &\rightarrow \text{V NP [Adv] [PP}_\text{M}\text{]} \\
\text{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down} \\
\text{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\text{PP}_\text{M} &\rightarrow \text{P}_\text{M}\text{ NP} \\
\text{P}_\text{M} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

147,123,874,800 sentences without recursion

$\emptyset$

Generating sentences

$$
\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{NP}\ \mathrm{VP} \\
\mathrm{NP} &\rightarrow \mathrm{D}\ [\mathrm{A}]\ \mathrm{N}\ [\mathrm{PP}] \\
\mathrm{D} &\rightarrow \textit{an} \mid \textit{the} \\
\mathrm{A} &\rightarrow \textit{blue} \mid \textit{red} \\
\mathrm{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object} \\
\mathrm{PP} &\rightarrow \mathrm{P}\ \mathrm{NP} \\
\mathrm{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\mathrm{VP} &\rightarrow \mathrm{V}\ \mathrm{NP}\ [\mathrm{Adv}]\ [\mathrm{PP_M}] \\
\mathrm{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down} \\
\mathrm{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\mathrm{PP_M} &\rightarrow \mathrm{P_M}\ \mathrm{NP} \\
\mathrm{P_M} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

147,123,874,800 sentences without recursion

“carried”

Generating sentences

$$
\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{NP}\ \mathrm{VP}\\
\mathrm{NP} &\rightarrow \mathrm{D}\ [\mathrm{A}]\ \mathrm{N}\ [\mathrm{PP}]\\
\mathrm{D} &\rightarrow \textit{an} \mid \textit{the}\\
\mathrm{A} &\rightarrow \textit{blue} \mid \textit{red}\\
\mathrm{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object}\\
\mathrm{PP} &\rightarrow \mathrm{P}\ \mathrm{NP}\\
\mathrm{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of}\\
\mathrm{VP} &\rightarrow \mathrm{V}\ \mathrm{NP}\ [\mathrm{Adv}]\ [\mathrm{PP_M}]\\
\mathrm{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down}\\
\mathrm{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly}\\
\mathrm{PP_M} &\rightarrow \mathrm{P_M}\ \mathrm{NP}\\
\mathrm{P_M} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

147,123,874,800 sentences without recursion

“the person carried”

Generating sentences

$$
\begin{aligned}
\mathrm{S} &\rightarrow \mathrm{NP}\ \mathrm{VP} \\
\mathrm{NP} &\rightarrow \mathrm{D}\ [\mathrm{A}]\ \mathrm{N}\ [\mathrm{PP}] \\
\mathrm{D} &\rightarrow \textit{an} \mid \textit{the} \\
\mathrm{A} &\rightarrow \textit{blue} \mid \textit{red} \\
\mathrm{N} &\rightarrow \textit{person} \mid \textit{backpack} \mid \textit{chair} \mid \textit{bin} \mid \textit{object} \\
\mathrm{PP} &\rightarrow \mathrm{P}\ \mathrm{NP} \\
\mathrm{P} &\rightarrow \textit{to the left of} \mid \textit{to the right of} \\
\mathrm{VP} &\rightarrow \mathrm{V}\ \mathrm{NP}\ [\mathrm{Adv}]\ [\mathrm{PP}_{\mathrm{M}}] \\
\mathrm{V} &\rightarrow \textit{approached} \mid \textit{carried} \mid \textit{picked up} \mid \textit{put down} \\
\mathrm{Adv} &\rightarrow \textit{quickly} \mid \textit{slowly} \\
\mathrm{PP}_{\mathrm{M}} &\rightarrow \mathrm{P}_{\mathrm{M}}\ \mathrm{NP} \\
\mathrm{P}_{\mathrm{M}} &\rightarrow \textit{towards} \mid \textit{away from}
\end{aligned}
$$

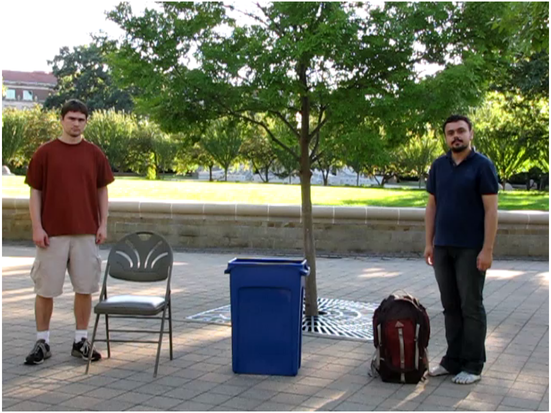
147,123,874,800 sentences without recursion

“the person carried the backpack”

Generated sentences

Generated sentences

Generated sentences



The person to the right of the bin picked up the backpack.

Recognition $P(\text{sentence}, \text{video})$ Narayanaswamy *et al.* 2014

Retrieval $\operatorname*{argmax}_{v \in V} P(s, v)$ Barret *et al.* 2016

Generation $\operatorname*{argmax}_{s \in L} P(s, v)$ Yu *et al.* 2015, Narayanaswamy *et al.* 2

Question answering

Disambiguation

Language acquisition

Follow commands

Paraphrasing

Translation

Common sense reasoning

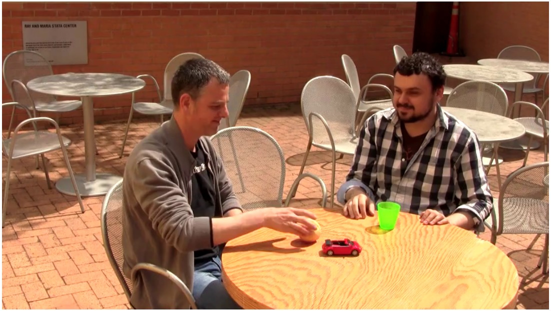
Planning

…

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation		
Language acquisition		
Follow commands		
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
…		

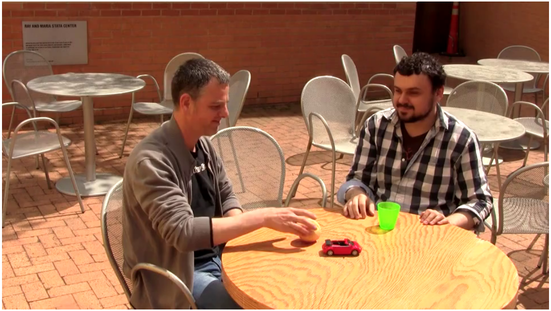
Question answering

Question answering



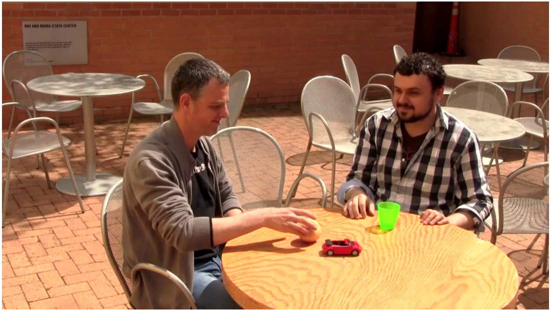
Question answering

Question answering



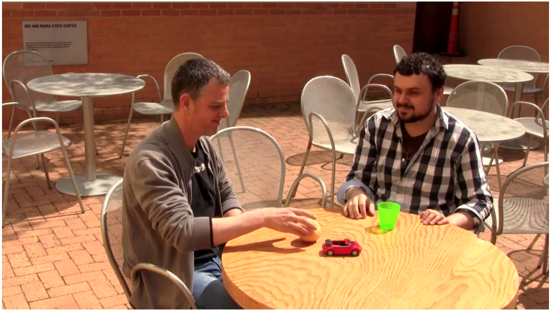
What did the person put on top of the red car?

Question answering



What did the person put on top of the red car?
The person put NP on top of the red car.

Question answering



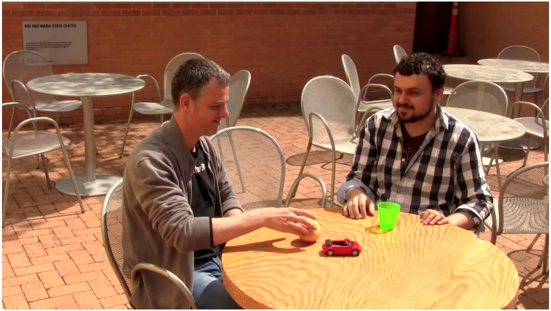
What did the person put on top of the red car?
The person put NP on top of the red car.
The person put the pear on top of the red car.

Question answering

Question answering

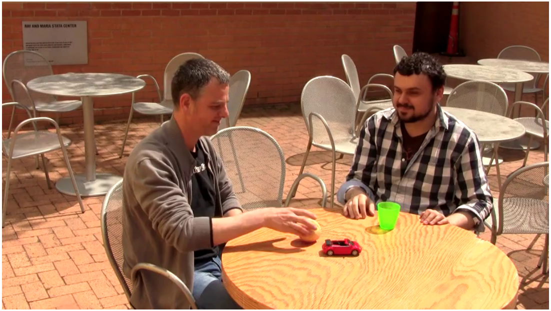
Question answering

Question answering



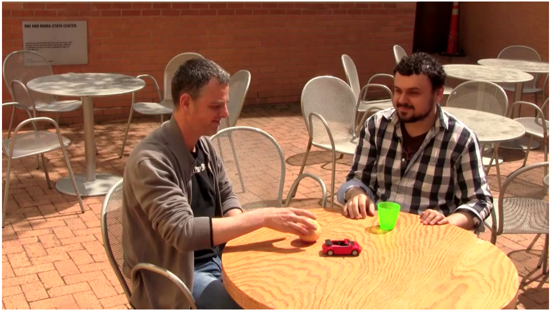
Who put an object on top of the red car?

Question answering with specificity



Who put an object on top of the red car?
NP put an object on top of the red car.

Question answering with specificity



Who put an object on top of the red car?

NP put an object on top of the red car.

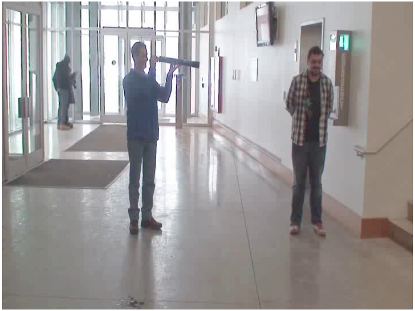
The person on the left of the car put an object on top of the red car.

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation		
Language acquisition		
Follow commands		
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
…		

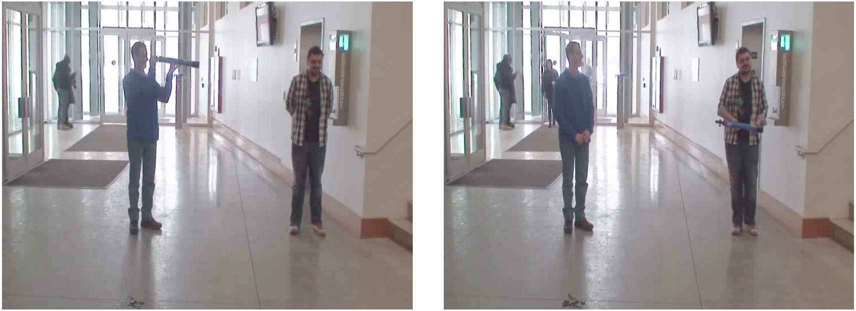
Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \text{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition		
Follow commands		
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
…		

I saw the man with the telescope.

I saw the man with the telescope.

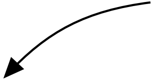


I saw the man with the telescope.

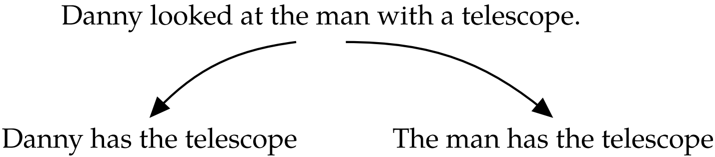


Danny looked at the man with a telescope.

Danny looked at the man with a telescope.



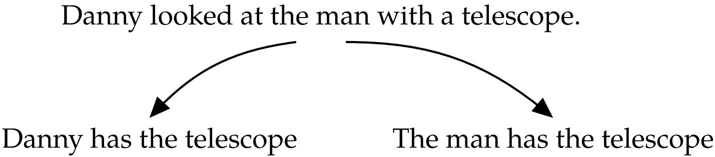
Danny has the telescope



Danny looked at the man with a telescope.

Danny has the telescope

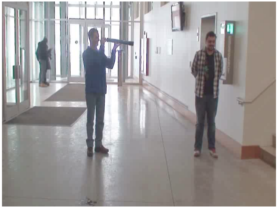
The man has the telescope

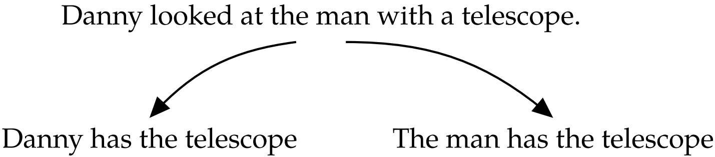


Danny looked at the man with a telescope.

Danny has the telescope

The man has the telescope

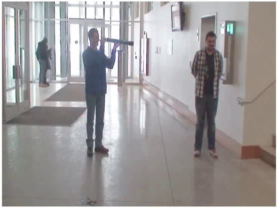


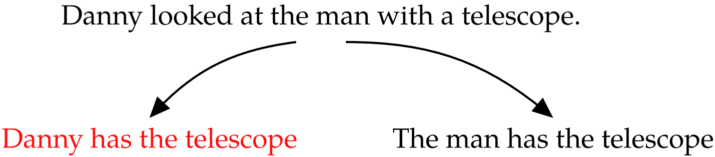


Danny looked at the man with a telescope.

Danny has the telescope

The man has the telescope

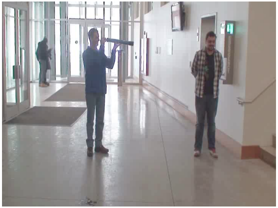




Danny looked at the man with a telescope.

Danny has the telescope

The man has the telescope

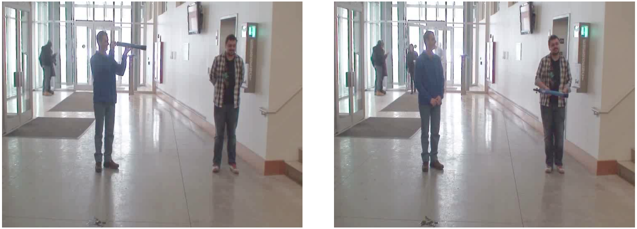


Ambiguities

Ambiguities

PP Attachment

Danny looked at the man with a telescope.

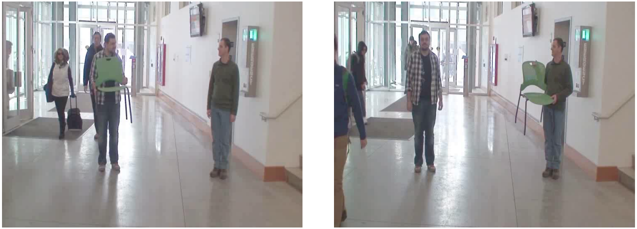


Ambiguities

PP Attachment

VP Attachment

Andrei approached the person holding a green chair.



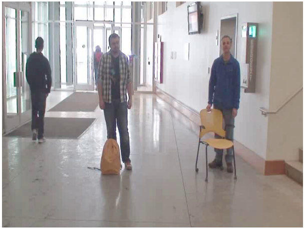
Ambiguities

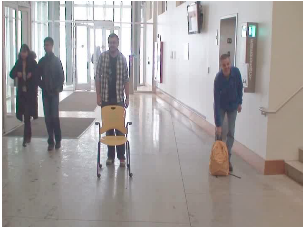
PP Attachment

VP Attachment

Conjunction

Danny and Andrei picked up the yellow bag and chair.





Ambiguities

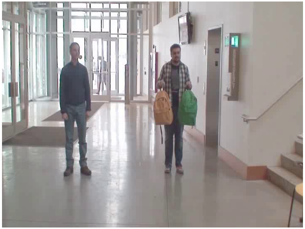
PP Attachment

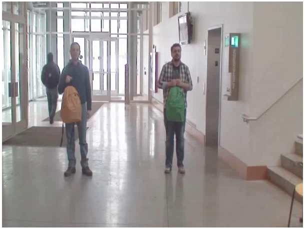
VP Attachment

Conjunction

Logical Form

Someone put down the bags.





Ambiguities

PP Attachment

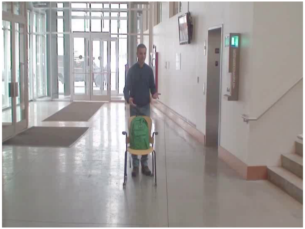
VP Attachment

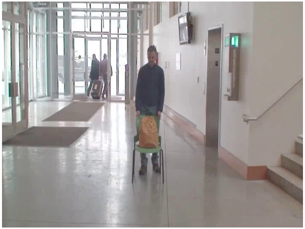
Conjunction

Logical Form

Anaphora

Danny picked up the bag and the chair. It is yellow.





Ambiguities

PP Attachment

VP Attachment

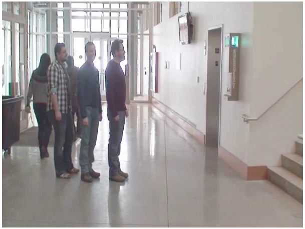
Conjunction

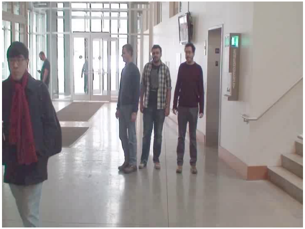
Logical Form

Anaphora

Ellipsis

Danny left Andrei. Also Yevgeni.





Danny and Andrei moved a chair.

Danny and Andrei moved a chair.

Danny and Andrei move the same chair.

Danny and Andrei moved a chair.

Danny and Andrei move the same chair.

$$\exists x\ \mathbf{chair}(x)$$

$$\mathbf{move}(\text{Danny}, x), \mathbf{move}(\text{Andrei}, x)$$

Danny and Andrei moved a chair.

Danny and Andrei move the same chair.

$$\exists x\ \mathbf{chair}(x)$$

$$\mathbf{move}(\text{Danny}, x), \mathbf{move}(\text{Andrei}, x)$$

Danny and Andrei move different chairs.

Danny and Andrei moved a chair.

Danny and Andrei move the same chair.

$$\exists x\ \mathbf{chair}(x)$$

$$\mathbf{move}(\text{Danny}, x), \mathbf{move}(\text{Andrei}, x)$$

Danny and Andrei move different chairs.

$$\exists xy\ \mathbf{chair}(x), \mathbf{chair}(y)$$

$$\mathbf{move}(\text{Danny}, x), \mathbf{move}(\text{Andrei}, y), x \neq y$$

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \text{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition		
Follow commands		
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
…		

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \mathrm{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\operatorname*{argmax}_{\theta} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015
Follow commands		
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
…		

Language learning

Language learning

Split into two variants:

Language learning

Split into two variants:

- Lexicon

Language learning

Split into two variants:

- Lexicon
- Syntax

Language learning

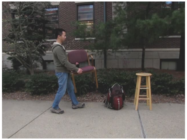
Split into two variants:

- Lexicon
- Syntax

Language learning: Lexicon

Yu *et al.* 2014

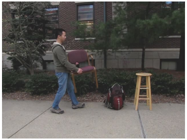
Language learning: Lexicon



Yu *et al.* 2014

Language learning: Lexicon

The person picked up the chair.

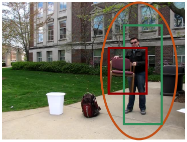


The chair approached the backpack.

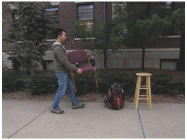
The person picked up the backpack.

Yu *et al.* 2014

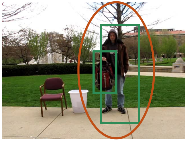
Language learning: Lexicon



The person picked up the chair.



The chair approached the backpack.

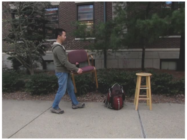


The person picked up the backpack.

Yu *et al.* 2014

Language learning: Lexicon

The person picked up the chair.

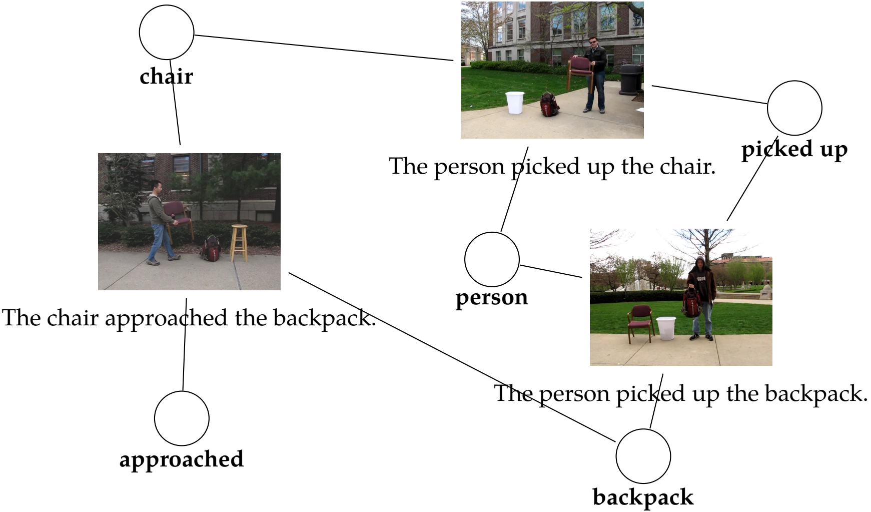


The chair approached the backpack.

The person picked up the backpack.

Yu *et al.* 2014

Language learning: Lexicon



chair

The person picked up the chair.

picked up

The chair approached the backpack.

person

The person picked up the backpack.

approached

backpack

Yu *et al.* 2014

Language learning

Split into two variants:

- Lexicon
- Syntax

Language learning

Split into two variants:

- Lexicon
- Syntax

Language learning: Syntax

Language learning: Syntax

Danny approached the chair with a bag.

Language learning: Syntax

Danny approached the chair with a bag.

↓

parser

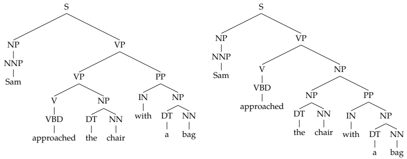
Language learning: Syntax

Danny approached the chair with a bag.

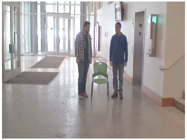
↓

parser

↓

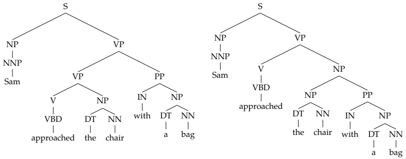


Language learning: Syntax

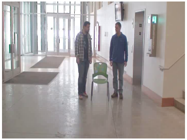


Danny approached the chair with a bag.

parser

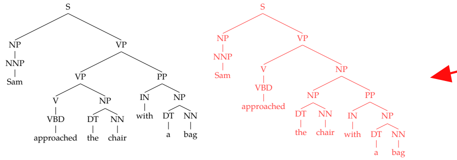


Language learning: Syntax



Danny approached the chair with a bag.

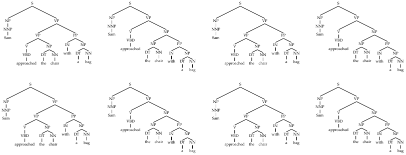
parser



Language learning: Syntax

Danny approached the chair with a bag.

parser

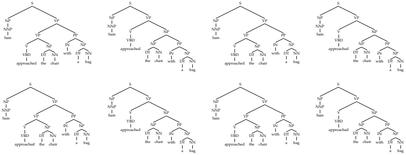


• • •

Language learning: Syntax

Danny approached the chair with a bag.

≈parser

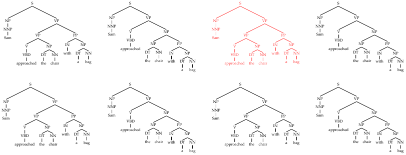

S NP VP NNP Sam VP PP V NP IN NP VBD DT NN with DT NN approached the chair a bag
S NP VP NNP Sam V NP VBD NP PP approached DT NN IN NP the chair with DT NN a bag
S NP VP NNP Sam VP PP V NP IN NP VBD DT NN with DT NN approached the chair a bag
S NP VP NNP Sam V NP VBD NP PP approached DT NN IN NP the chair with DT NN a bag
S NP VP NNP Sam VP PP V NP IN NP VBD DT NN with DT NN approached the chair a bag
S NP VP NNP Sam V NP VBD NP PP approached DT NN IN NP the chair with DT NN a bag
S NP VP NNP Sam VP PP V NP IN NP VBD DT NN with DT NN approached the chair a bag
S NP VP NNP Sam V NP VBD NP PP approached DT NN IN NP the chair with DT NN a bag

• • •

Language learning: Syntax

Danny approached the chair with a bag.

≈parser



• • •

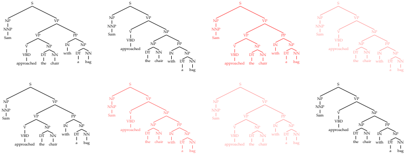
Language learning: Syntax

Danny approached the chair with a bag.

↓

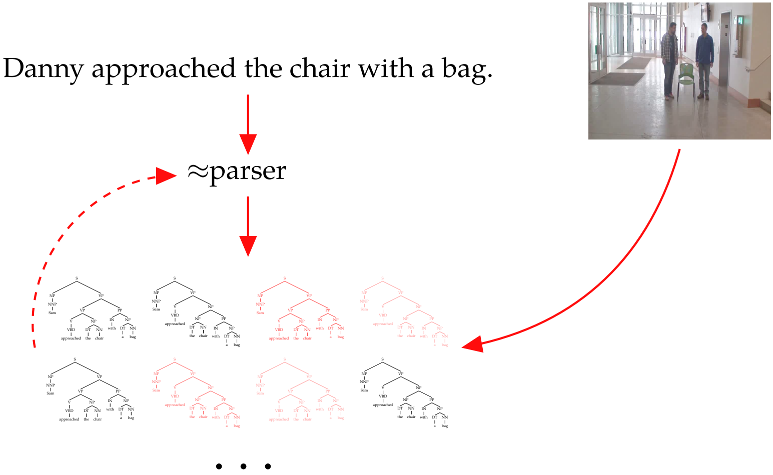
≈parser

↓


S NP VP NNP Sam VP PP V NP IN NP VBD DT NN with DT NN approached the chair a bag
S NP VP NNP Sam V NP VBD NP PP approached DT NN IN NP the chair with DT NN a bag
S NP VP NNP Sam VP PP V NP IN NP VBD DT NN with DT NN approached the chair a bag
S NP VP NNP Sam V NP VBD NP PP approached DT NN IN NP the chair with DT NN a bag
S NP VP NNP Sam VP PP V NP IN NP VBD DT NN with DT NN approached the chair a bag
S NP VP NNP Sam V NP VBD NP PP approached DT NN IN NP the chair with DT NN a bag
S NP VP NNP Sam VP PP V NP IN NP VBD DT NN with DT NN approached the chair a bag
S NP VP NNP Sam V NP VBD NP PP approached DT NN IN NP the chair with DT NN a bag

• • •

Language learning: Syntax



Danny approached the chair with a bag.

≈parser

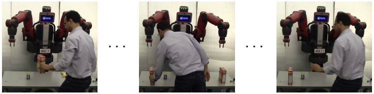
Pilley and Reid 2011

Pilley and Reid 2011

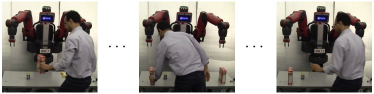
Pilley and Reid 2011

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \text{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\operatorname*{argmax}_{\theta} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015
Follow commands		
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
…		

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\mathop{\mathrm{argmax}}\limits_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\mathop{\mathrm{argmax}}\limits_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\mathop{\mathrm{argmax}}\limits_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\mathop{\mathrm{argmax}}\limits_{i \in \mathrm{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\mathop{\mathrm{argmax}}\limits_{\theta} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015
Follow commands	$\mathop{\mathrm{argmax}}\limits_{p} \int_{v^+} P(C(s), v^+ v)\, E(v^+, p, v)$	Paul *et al.* 2017
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
…		



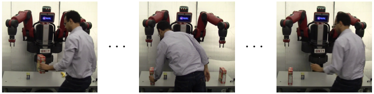
The box I put down on the table is expensive.



The box I put down on the table is expensive.

$\text{Human}(x) \wedge \text{PutDown}(x, y) \wedge \text{Box}(y) \wedge \text{On}(y, z) \wedge \text{Table}(z)$

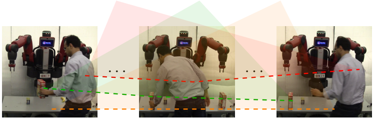
$\text{Assert}(\text{Expensive}(y))$

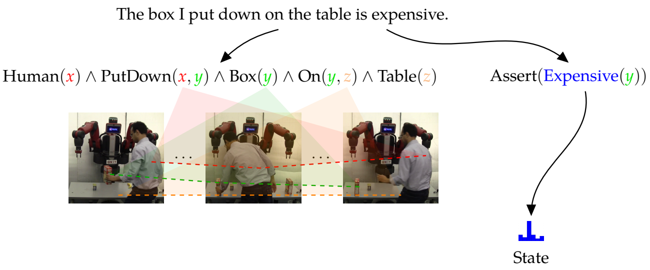


The box I put down on the table is expensive.

$\text{Human}(x) \wedge \text{PutDown}(x, y) \wedge \text{Box}(y) \wedge \text{On}(y, z) \wedge \text{Table}(z)$

$\text{Assert}(\text{Expensive}(y))$



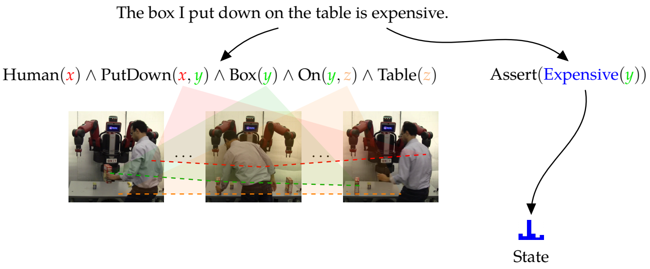


The box I put down on the table is expensive.

$\text{Human}(x) \wedge \text{PutDown}(x, y) \wedge \text{Box}(y) \wedge \text{On}(y, z) \wedge \text{Table}(z)$

$\text{Assert}(\text{Expensive}(y))$

State



The box I put down on the table is expensive.

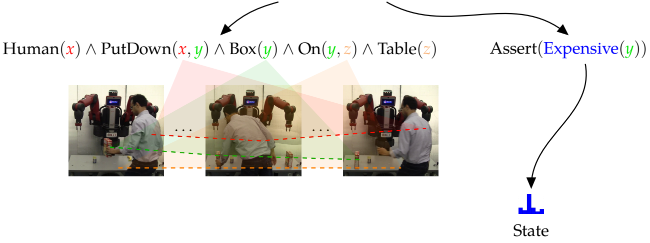
$\text{Human}(x) \wedge \text{PutDown}(x, y) \wedge \text{Box}(y) \wedge \text{On}(y, z) \wedge \text{Table}(z)$

$\text{Assert}(\text{Expensive}(y))$

State

Pick up the expensive box someone put down next to the can.

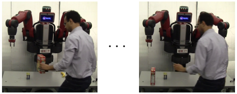
The box I put down on the table is expensive.



$\mathrm{Human}(x) \wedge \mathrm{PutDown}(x, y) \wedge \mathrm{Box}(y) \wedge \mathrm{On}(y, z) \wedge \mathrm{Table}(z)$

$\mathrm{Assert}(\mathrm{Expensive}(y))$

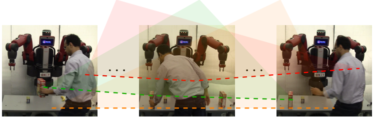
Pick up the expensive box someone put down next to the can.



The box I put down on the table is expensive.

$\text{Human}(x) \land \text{PutDown}(x, y) \land \text{Box}(y) \land \text{On}(y, z) \land \text{Table}(z)$

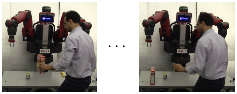
$\text{Assert}(\text{Expensive}(y))$



State

Pick up the expensive box someone put down next to the can.

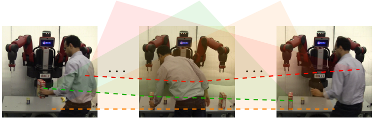
$\text{PutDown}(x, y) \land \text{Expensive}(y) \land \text{Can}(z) \land \text{Box}(y) \land \text{NextTo}(y, z)$



The box I put down on the table is expensive.

$\text{Human}(x) \wedge \text{PutDown}(x, y) \wedge \text{Box}(y) \wedge \text{On}(y, z) \wedge \text{Table}(z)$

$\text{Assert}(\text{Expensive}(y))$

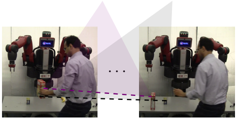


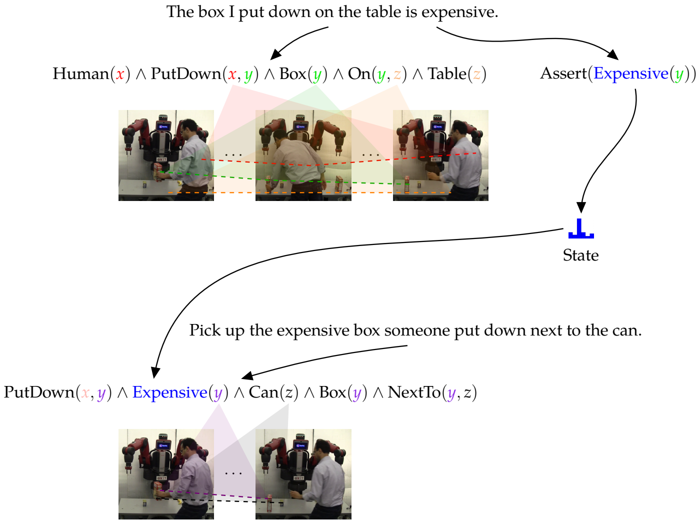
⊥

State

Pick up the expensive box someone put down next to the can.

$\text{PutDown}(x, y) \wedge \text{Expensive}(y) \wedge \text{Can}(z) \wedge \text{Box}(y) \wedge \text{NextTo}(y, z)$





The box I put down on the table is expensive.

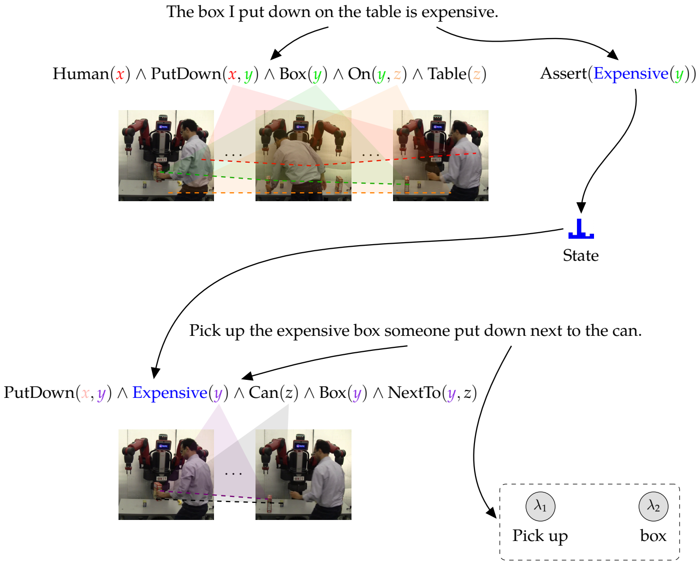
$\text{Human}(x) \wedge \text{PutDown}(x, y) \wedge \text{Box}(y) \wedge \text{On}(y, z) \wedge \text{Table}(z)$

$\text{Assert}(\text{Expensive}(y))$

State

Pick up the expensive box someone put down next to the can.

$\text{PutDown}(x, y) \wedge \text{Expensive}(y) \wedge \text{Can}(z) \wedge \text{Box}(y) \wedge \text{NextTo}(y, z)$



The box I put down on the table is expensive.

$\text{Human}(x) \wedge \text{PutDown}(x, y) \wedge \text{Box}(y) \wedge \text{On}(y, z) \wedge \text{Table}(z)$

$\text{Assert}(\text{Expensive}(y))$

⊥

State

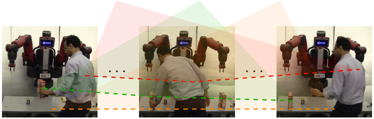
Pick up the expensive box someone put down next to the can.

$\text{PutDown}(x, y) \wedge \text{Expensive}(y) \wedge \text{Can}(z) \wedge \text{Box}(y) \wedge \text{NextTo}(y, z)$

The box I put down on the table is expensive.

$\text{Human}(x) \wedge \text{PutDown}(x, y) \wedge \text{Box}(y) \wedge \text{On}(y, z) \wedge \text{Table}(z)$

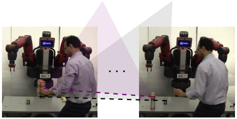
$\text{Assert}(\text{Expensive}(y))$

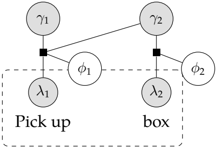


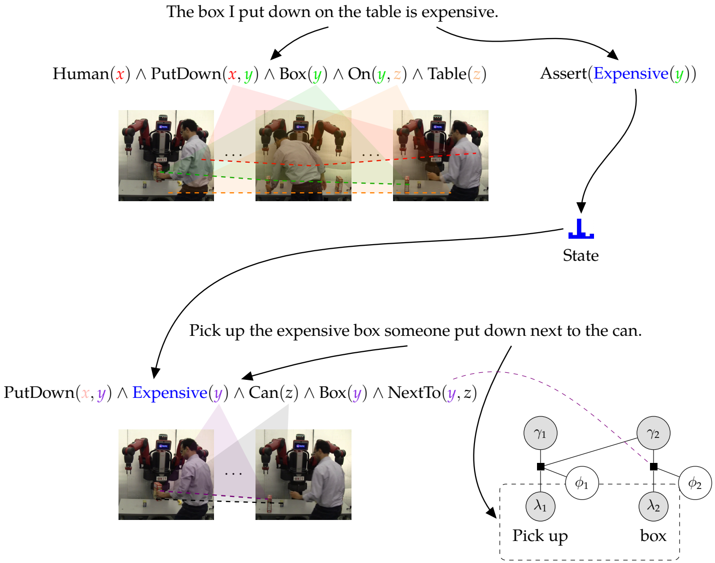
State

Pick up the expensive box someone put down next to the can.

$\text{PutDown}(x, y) \wedge \text{Expensive}(y) \wedge \text{Can}(z) \wedge \text{Box}(y) \wedge \text{NextTo}(y, z)$







The box I put down on the table is expensive.

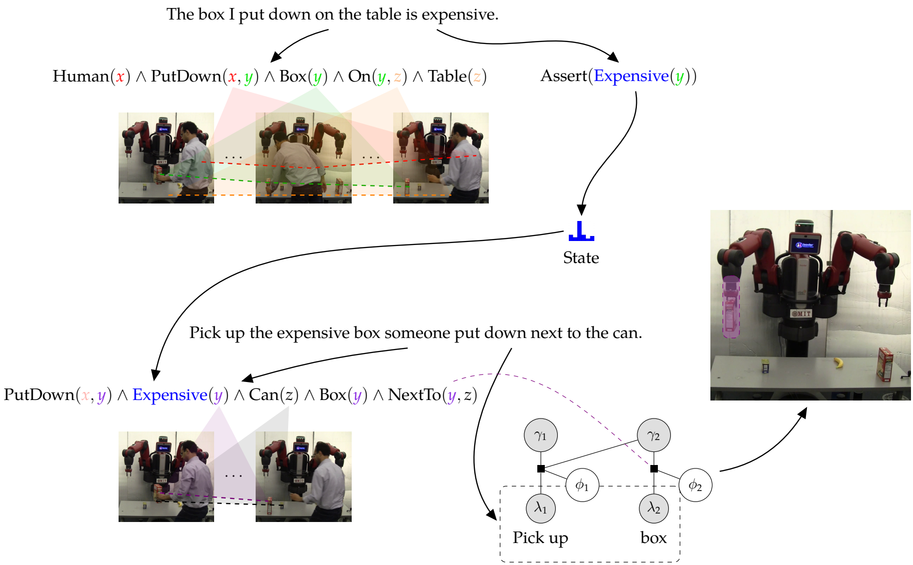
$\text{Human}(x) \wedge \text{PutDown}(x, y) \wedge \text{Box}(y) \wedge \text{On}(y, z) \wedge \text{Table}(z)$

$\text{Assert}(\text{Expensive}(y))$

State

Pick up the expensive box someone put down next to the can.

$\text{PutDown}(x, y) \wedge \text{Expensive}(y) \wedge \text{Can}(z) \wedge \text{Box}(y) \wedge \text{NextTo}(y, z)$



The box I put down on the table is expensive.

$\text{Human}(x) \wedge \text{PutDown}(x, y) \wedge \text{Box}(y) \wedge \text{On}(y, z) \wedge \text{Table}(z)$

$\text{Assert}(\text{Expensive}(y))$

$\perp$

State

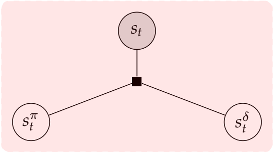
Pick up the expensive box someone put down next to the can.

$\text{PutDown}(x, y) \wedge \text{Expensive}(y) \wedge \text{Can}(z) \wedge \text{Box}(y) \wedge \text{NextTo}(y, z)$

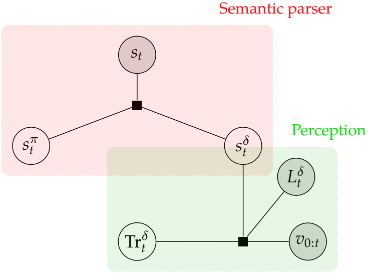
Understanding and following commands

Understanding and following commands

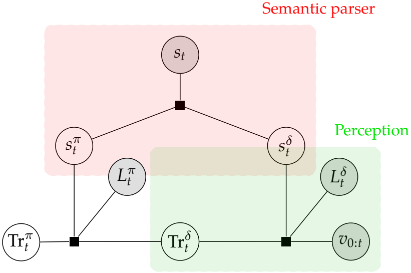
Semantic parser



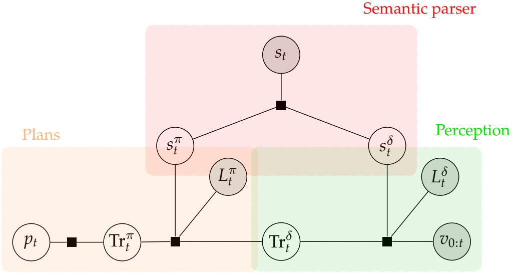
Understanding and following commands



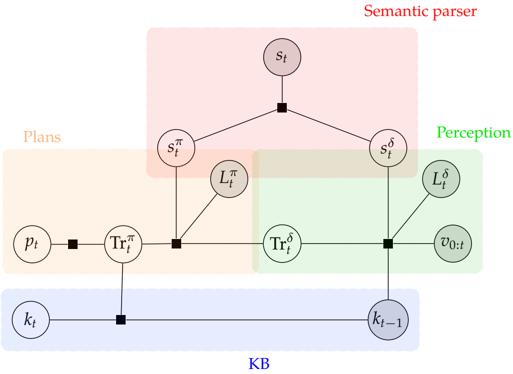
Understanding and following commands

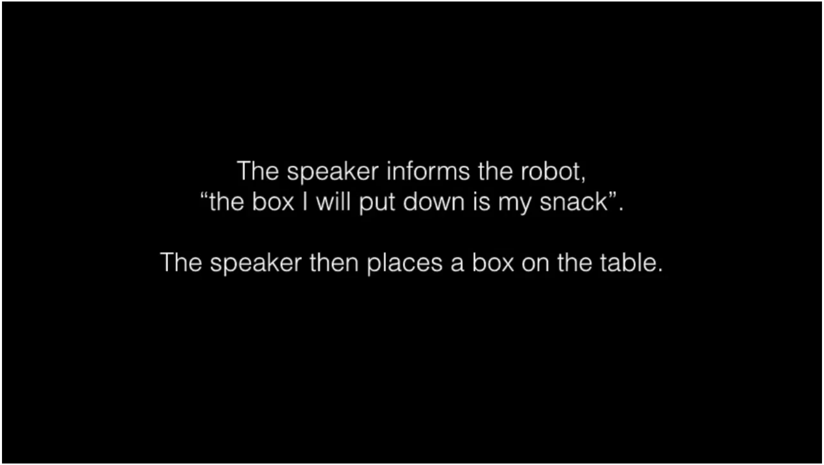


Understanding and following commands



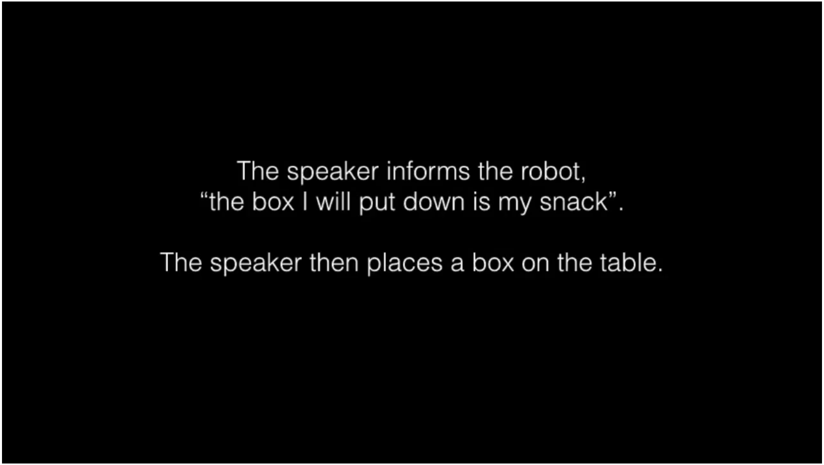
Understanding and following commands





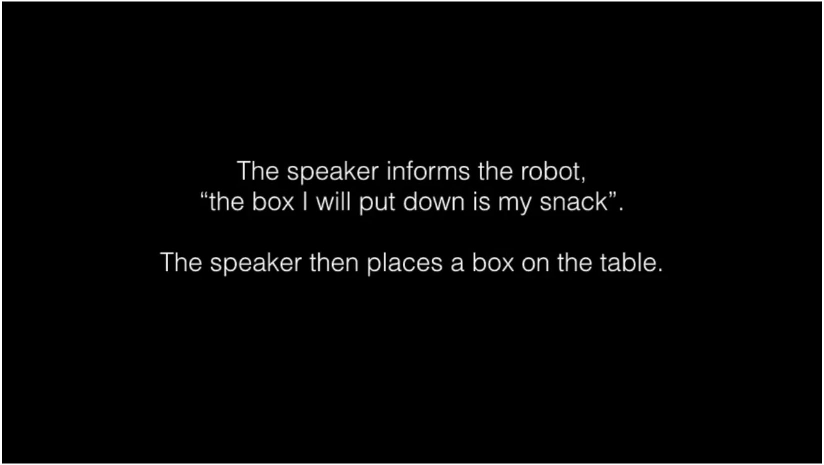
The speaker informs the robot,
“the box I will put down is my snack”.

The speaker then places a box on the table.



The speaker informs the robot,
“the box I will put down is my snack”.

The speaker then places a box on the table.



The speaker informs the robot,
"the box I will put down is my snack".

The speaker then places a box on the table.

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \text{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\operatorname*{argmax}_{\theta} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015
Follow commands	$\operatorname*{argmax}_{p} \int_{v^+} P(C(s), v^+ v)\, E(v^+, p, v)$	Paul *et al.* 2017
Paraphrasing		
Translation		
Common sense reasoning		
Planning		
…		

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \text{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\operatorname*{argmax}_{\theta} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015
Follow commands	$\operatorname*{argmax}_{p} \int_{v^+} P(C(s), v^+ v)\, E(v^+, p, v)$	Paul *et al.* 2017
Paraphrasing	$\int_{v} \lvert P(s, v) - P(s', v) \rvert$	
Translation		
Common sense reasoning		
Planning		
…		

Paraphrasing

Paraphrasing

The dark haired man is picking an object up from the floor.

The guy in the plaid shirt is lifting the yellow chair.

Paraphrasing

The dark haired man is picking an object up from the floor.

The guy in the plaid shirt is lifting the yellow chair.

The man with the chair walks away from someone.

The man walks away from someone with the chair.

Paraphrasing

Paraphrasing

The tall man gave the woman the red box.

Paraphrasing

The tall man gave the woman the red box.

$\overset{?}{\Rightarrow}$ The woman received the crimson box from the man.

Paraphrasing

The tall man gave the woman the red box.

$\overset{?}{\Rightarrow}$ The woman received the crimson box from the man.
$\overset{?}{\Rightarrow}$ The box was passed on by the body guard.

Paraphrasing is hard

The tall man gave the woman the red box.

$\overset{?}{\Rightarrow}$ The woman received the crimson box from the man.

$\overset{?}{\Rightarrow}$ The box was passed on by the body guard.

$\overset{?}{\Rightarrow}$ The box changed hands from the man to the woman.

Paraphrasing is hard

The tall man gave the woman the red box.

$\overset{?}{\Rightarrow}$ The woman received the crimson box from the man.

$\overset{?}{\Rightarrow}$ The box was passed on by the body guard.

$\overset{?}{\Rightarrow}$ The box changed hands from the man to the woman.

$\overset{?}{\Rightarrow}$ The man entered a room with a box. A woman left holding it.

Paraphrasing is hard

The tall man gave the woman the red box.

$\overset{?}{\Rightarrow}$ The woman received the crimson box from the man.

$\overset{?}{\Rightarrow}$ The box was passed on by the body guard.

$\overset{?}{\Rightarrow}$ The box changed hands from the man to the woman.

$\overset{?}{\Rightarrow}$ The man entered a room with a box. A woman left holding it.

$\overset{?}{\Rightarrow}$ The cat yawned.

Paraphrasing is hard

The tall man gave the woman the red box.

$\overset{?}{\Rightarrow}$ The woman received the crimson box from the man.

$\overset{?}{\Rightarrow}$ The box was passed on by the body guard.

$\overset{?}{\Rightarrow}$ The box changed hands from the man to the woman.

$\overset{?}{\Rightarrow}$ The man entered a room with a box. A woman left holding it.

$\overset{?}{\Rightarrow}$ The cat yawned.

$\overset{?}{\Rightarrow}$ The guard took the box from the woman.

Paraphrasing is hard

The tall man gave the woman the red box.

$\overset{?}{\Rightarrow}$ The woman received the crimson box from the man.

$\overset{?}{\Rightarrow}$ The box was passed on by the body guard.

$\overset{?}{\Rightarrow}$ The box changed hands from the man to the woman.

$\overset{?}{\Rightarrow}$ The man entered a room with a box. A woman left holding it.

$\overset{?}{\Rightarrow}$ The cat yawned.

$\overset{?}{\Rightarrow}$ The guard took the box from the woman.

$\overset{?}{\Rightarrow}$ The woman left the box behind.

Paraphrasing is hard

The tall man gave the woman the red box.

$\overset{?}{\Rightarrow}$ The woman received the crimson box from the man.

$\overset{?}{\Rightarrow}$ The box was passed on by the body guard.

$\overset{?}{\Rightarrow}$ The box changed hands from the man to the woman.

$\overset{?}{\Rightarrow}$ The man entered a room with a box. A woman left holding it.

$\overset{?}{\Rightarrow}$ The cat yawned.

$\overset{?}{\Rightarrow}$ The guard took the box from the woman.

$\overset{?}{\Rightarrow}$ The woman left the box behind.

$\overset{?}{\Rightarrow}$ The box was passed to the man.

Paraphrasing is hard

The tall man gave the woman the red box.

$\overset{?}{\Rightarrow}$ The woman received the crimson box from the man.

$\overset{?}{\Rightarrow}$ The box was passed on by the body guard.

$\overset{?}{\Rightarrow}$ The box changed hands from the man to the woman.

$\overset{?}{\Rightarrow}$ The man entered a room with a box. A woman left holding it.

$\overset{?}{\Rightarrow}$ The cat yawned.

$\overset{?}{\Rightarrow}$ The guard took the box from the woman.

$\overset{?}{\Rightarrow}$ The woman left the box behind.

$\overset{?}{\Rightarrow}$ The box was passed to the man.

$\overset{?}{\Rightarrow}$ The crate was on the floor while the woman picked up the dog.

Paraphrasing is hard

The tall man gave the woman the red box.

$\overset{?}{\Rightarrow}$ The woman received the crimson box from the man.

$\overset{?}{\Rightarrow}$ The box was passed on by the body guard.

$\overset{?}{\Rightarrow}$ The box changed hands from the man to the woman.

$\overset{?}{\Rightarrow}$ The man entered a room with a box. A woman left holding it.

$\overset{?}{\Rightarrow}$ The cat yawned.

$\overset{?}{\Rightarrow}$ The guard took the box from the woman.

$\overset{?}{\Rightarrow}$ The woman left the box behind.

$\overset{?}{\Rightarrow}$ The box was passed to the man.

$\overset{?}{\Rightarrow}$ The crate was on the floor while the woman picked up the dog.

$\overset{?}{\Rightarrow}$ The dog was on the floor while the woman picked up the crate.

Paraphrasing is hard

The tall man gave the woman the red box.

$\overset{?}{\Rightarrow}$ The woman received the crimson box from the man.

$\overset{?}{\Rightarrow}$ The box was passed on by the body guard.

$\overset{?}{\Rightarrow}$ The box changed hands from the man to the woman.

$\overset{?}{\Rightarrow}$ The man entered a room with a box. A woman left holding it.

$\overset{?}{\Rightarrow}$ The cat yawned.

$\overset{?}{\Rightarrow}$ The guard took the box from the woman.

$\overset{?}{\Rightarrow}$ The woman left the box behind.

$\overset{?}{\Rightarrow}$ The box was passed to the man.

$\overset{?}{\Rightarrow}$ The crate was on the floor while the woman picked up the dog.

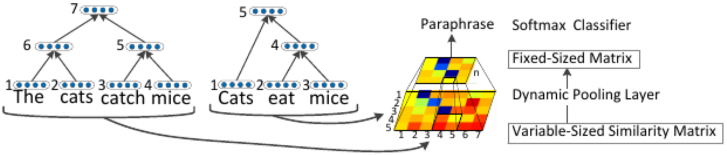
$\overset{?}{\Rightarrow}$ The dog was on the floor while the woman picked up the crate.

…

Paraphrasing today

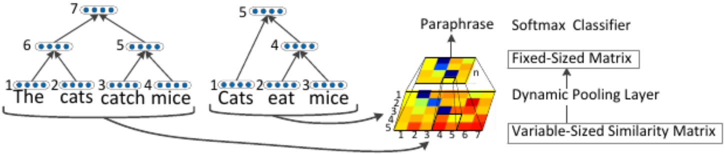
Socher *et al.* 2011, Cheng and Kartsaklis 2015

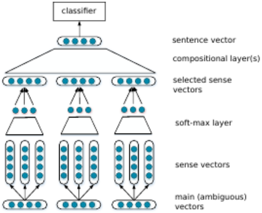
Paraphrasing today



Socher *et al.* 2011, Cheng and Kartsaklis 2015

Paraphrasing today





Socher *et al.* 2011, Cheng and Kartsaklis 2015

Paraphrasing with vision

Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

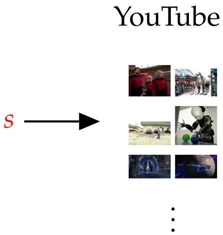
Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

s

Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

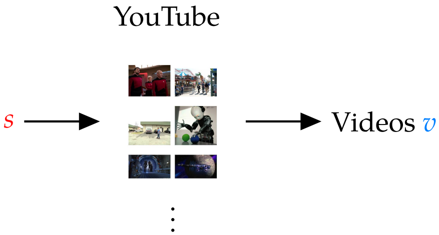


YouTube

s

Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

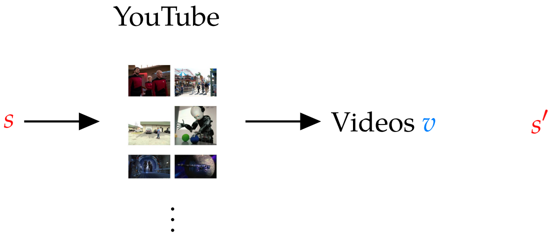


YouTube

s → ⋮ → Videos v

Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

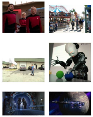


Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

YouTube

$s \longrightarrow$



$\vdots$

$\longrightarrow$ Videos v

s'

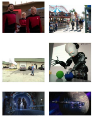
$$\underbrace{\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad}_{\displaystyle\sum_{v} |S(s,v) - S(s',v)|^2}$$

Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

YouTube

$s \longrightarrow$



$\longrightarrow$ Videos v

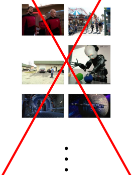
s'

$\vdots$

$$\underbrace{\hspace{20em}}_{\sum_{v} |S(s,v) - S(s',v)|^2}$$

Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$



$$\underbrace{s \longrightarrow \cancel{\text{YouTube}} \longrightarrow \text{Videos } v \qquad s'}_{\sum_{v} |S(s,v)-S(s',v)|^2}$$

Paraphrasing with vision

$$s \overset{?}{\Rightarrow} s'$$

$$\underbrace{s \xrightarrow{\text{sample}} \text{Videos } v \qquad\qquad s'}_{\int_v |S(s,v)-S(s',v)|^2}$$

Paraphrasing with imagination

$$s \overset{?}{\Rightarrow} s'$$

$$\underbrace{s \xrightarrow{\text{sample}} \text{Videos } v \qquad\qquad s'}_{\int_v |S(s,v) - S(s',v)|^2}$$

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\underset{v\in V}{\operatorname{argmax}}\ P(s,v)$	Barret *et al.* 2016
Generation	$\underset{s\in L}{\operatorname{argmax}}\ P(s,v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\underset{s\in L}{\operatorname{argmax}}\ P(\mathrm{Q}(s,s_q),v)$	Barbu *et al.* in prep.
Disambiguation	$\underset{i\in \text{parser}(s)}{\operatorname{argmax}}\ P(i,v)$	Berzak *et al.* 2015
Language acquisition	$\underset{\theta}{\operatorname{argmax}} \prod_{s,v} P(s(\theta),v)$	Yu *et al.* 2015
Follow commands	$\underset{p}{\operatorname{argmax}} \int_{v^+} P(C(s),v^+v)\, E(v^+,p,v)$	Paul *et al.* 2017
Paraphrasing	$\int_v \lvert P(s,v) - P(s',v) \rvert$	
Translation		
Common sense reasoning		
Planning		
…		

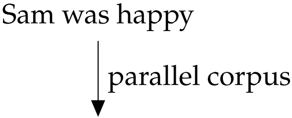
Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \text{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\operatorname*{argmax}_{\theta} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015
Follow commands	$\operatorname*{argmax}_{p} \int_{v^+} P(C(s), v^+ v)\, E(v^+, p, v)$	Paul *et al.* 2017
Paraphrasing	$\int_{v} \lvert P(s, v) - P(s', v) \rvert$	
Translation	$\operatorname*{argmin}_{s' \in L'} \int_{v} \lvert P(s, v) - P(s', v) \rvert$	
Common sense reasoning		
Planning		
…		

Statistical machine translation

Statistical machine translation

Sam was happy

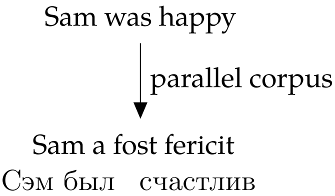
Statistical machine translation



Sam was happy

parallel corpus

Statistical machine translation



Sam was happy

parallel corpus

Sam a fost fericit

Сэм был счастлив

Statistical machine translation

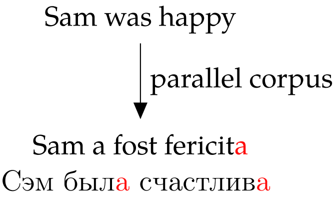
Sam was happy

↓ parallel corpus

Sam a fost fericita

Сэм была счастлива

Statistical machine translation



Sam was happy

parallel corpus

Sam a fost fericita

Сэм была счастлива

In Thai you specify siblings by age not gender.

Statistical machine translation

Sam was happy

↓ parallel corpus

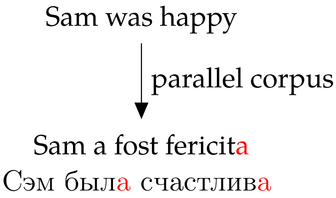
Sam a fost fericita

Сэм была счастлива

In Thai you specify siblings by age not gender.

In English you specify relative time but you don't need to in Chinese.

Statistical machine translation



Sam was happy

parallel corpus

Sam a fost fericita

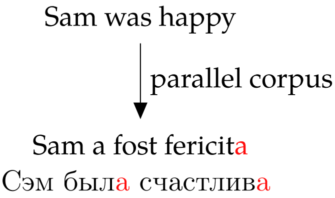
Сэм была счастлива

In Thai you specify siblings by age not gender.

In English you specify relative time but you don’t need to in Chinese.

Guugu Yimithirr language only uses absolute directions.

Statistical machine translation



Sam was happy

parallel corpus

Sam a fost fericita

Сэм была счастлива

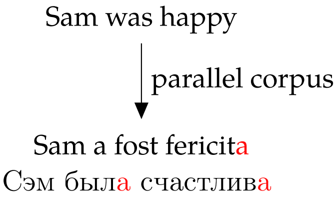
In Thai you specify siblings by age not gender.

In English you specify relative time but you don’t need to in Chinese.

Guugu Yimithirr language only uses absolute directions.

Many languages don’t distinguish blue/green.

Statistical machine translation



Sam was happy

parallel corpus

Sam a fost fericita

Сэм была счастлива

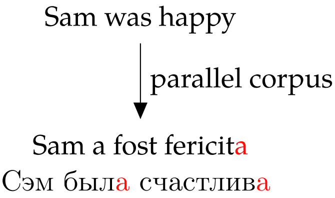
In Thai you specify siblings by age not gender.

In English you specify relative time but you don’t need to in Chinese.

Guugu Yimithirr language only uses absolute directions.

Many languages don’t distinguish blue/green.
Swahili specifies color as “the color of X“.

Statistical machine translation



Sam was happy

parallel corpus

Sam a fost fericita

Сэм была счастлива

In Thai you specify siblings by age not gender.

In English you specify relative time but you don't need to in Chinese.

Guugu Yimithirr language only uses absolute directions.

Many languages don't distinguish blue/green.
Swahili specifies color as "the color of *X*".

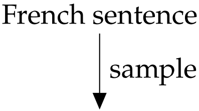
In Turkish you have to report if something is hearsay.

Translation by imagination

Translation by imagination

French sentence

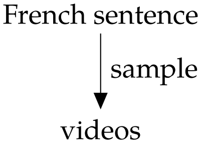
Translation by imagination



French sentence

sample

Translation by imagination

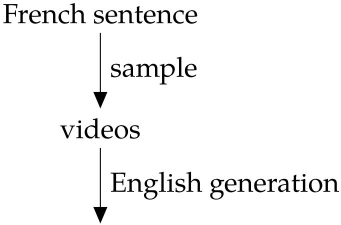


French sentence

sample

videos

Translation by imagination



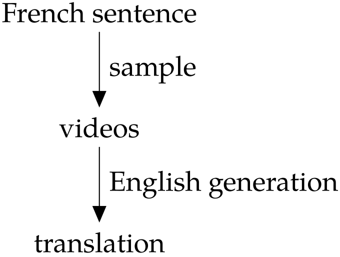
French sentence

sample

videos

English generation

Translation by imagination



French sentence

sample

videos

English generation

translation

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\text{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \text{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\operatorname*{argmax}_{\theta} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015
Follow commands	$\operatorname*{argmax}_{p} \int_{v+} P(C(s), v^{+} v)\, E(v^{+}, p, v)$	Paul *et al.* 2017
Paraphrasing	$\int_{v} \lvert P(s, v) - P(s', v) \rvert$	
Translation	$\operatorname*{argmin}_{s' \in L'} \int_{v} \lvert P(s, v) - P(s', v) \rvert$	
Common sense reasoning		
Planning		
…		

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \text{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\operatorname*{argmax}_{\theta} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015
Follow commands	$\operatorname*{argmax}_{p} \int_{v^+} P(\mathrm{C}(s), v^+ v)\, E(v^+, p, v)$	Paul *et al.* 2017
Paraphrasing	$\int_{v} \lvert P(s, v) - P(s', v) \rvert$	
Translation	$\operatorname*{argmin}_{s' \in L'} \int_{v} \lvert P(s, v) - P(s', v) \rvert$	
Common sense reasoning	$\operatorname*{argmax}_{s \in L} \int_{v} P(s_q, v) \; P(\mathrm{Q}(s, s_q), v)$	
Planning		
…		

Common sense reasoning

The trophy doesn't fit on the shelf.

Levesque *et al.* 2011

Common sense reasoning

The trophy doesn’t fit on the shelf because it’s too small.

Levesque *et al.* 2011

Common sense reasoning

The trophy doesn't fit on the shelf because it's too small.

Levesque *et al.* 2011

Common sense reasoning

The trophy doesn't fit on the shelf because it's too small.

Levesque *et al.* 2011

Common sense reasoning

The trophy doesn't fit on the shelf because it's too small.

Levesque *et al.* 2011

Common sense reasoning

The trophy doesn't fit on the shelf because it's too small.
The trophy doesn't fit on the shelf because it's too large.

Levesque *et al.* 2011

Common sense reasoning

The trophy doesn't fit on the shelf because it's too small.
The trophy doesn't fit on the shelf because it's too large.

Levesque *et al.* 2011

Common sense reasoning

The trophy doesn’t fit on the shelf because it’s too small.
The trophy doesn’t fit on the shelf because it’s too large.

Levesque *et al.* 2011

Common sense reasoning

The trophy doesn’t fit on the shelf because it’s too small.
The trophy doesn’t fit on the shelf because it’s too large.

Levesque *et al.* 2011

Common sense reasoning with vision

Common sense reasoning with vision

$$\frac{s}{s'}$$

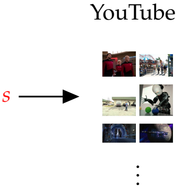
Common sense reasoning with vision

$$\frac{s}{s'}$$

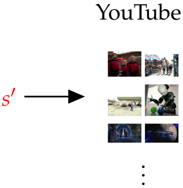
s s'

Common sense reasoning with vision

$$\frac{s}{s'}$$


YouTube

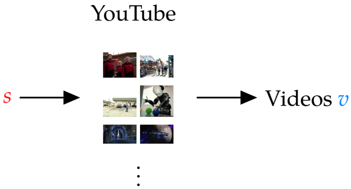
$s \longrightarrow$


YouTube

$s' \longrightarrow$

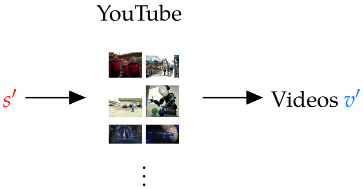
Common sense reasoning with vision

$$\frac{s}{s'}$$


YouTube

$s \longrightarrow$ $\longrightarrow$ Videos v

⋮

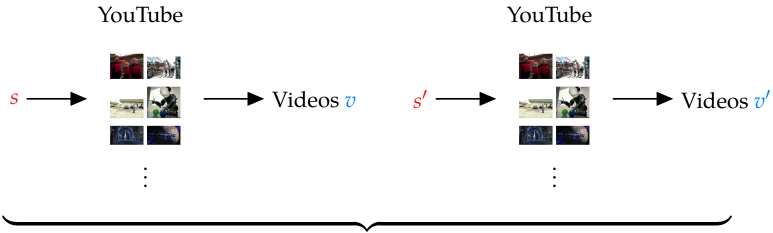

YouTube

$s' \longrightarrow$ $\longrightarrow$ Videos v'

⋮

Common sense reasoning with vision

$$\frac{s}{s'}$$



YouTube

$s \longrightarrow$ $\longrightarrow$ Videos v

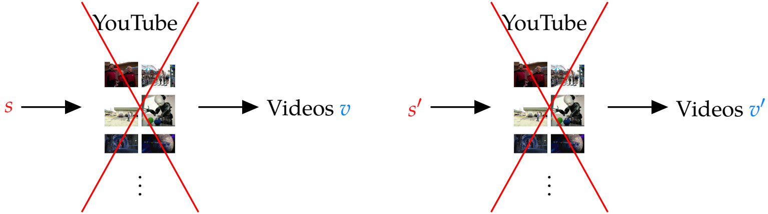
YouTube

$s' \longrightarrow$ $\longrightarrow$ Videos v'

$$\frac{\sum_{v} S(s, v)}{\sum_{v'} S(s', v')}$$

Common sense reasoning with vision

$$\frac{s}{s'}$$



$s \longrightarrow$ YouTube $\longrightarrow$ Videos v

$s' \longrightarrow$ YouTube $\longrightarrow$ Videos v'

$$\frac{\sum_{v} S(s, v)}{\sum_{v'} S(s', v')}$$

Common sense reasoning with vision

$$\frac{s}{s'}$$

$$\underbrace{s \xrightarrow{\text{sample}} \text{Videos } v \qquad s' \xrightarrow{\text{sample}} \text{Videos } v'}_{\displaystyle \frac{\int_{v} S(s, v)}{\int_{v'} S(s', v')}}$$

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \text{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\operatorname*{argmax}_{\theta} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015
Follow commands	$\operatorname*{argmax}_{p} \int_{v^+} P(\mathrm{C}(s), v^+ v)\, E(v^+, p, v)$	Paul *et al.* 2017
Paraphrasing	$\int_{v} \lvert P(s, v) - P(s', v) \rvert$	
Translation	$\operatorname*{argmin}_{s' \in L'} \int_{v} \lvert P(s, v) - P(s', v) \rvert$	
Common sense reasoning	$\operatorname*{argmax}_{s \in L} \int_{v} P(s_q, v)\ P(\mathrm{Q}(s, s_q), v)$	
Planning		

…

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\underset{v \in V}{\operatorname{argmax}}\ P(s, v)$	Barret *et al.* 2016
Generation	$\underset{s \in L}{\operatorname{argmax}}\ P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\underset{s \in L}{\operatorname{argmax}}\ P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\underset{i \in \mathrm{parser}(s)}{\operatorname{argmax}}\ P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\underset{\theta}{\operatorname{argmax}} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015
Follow commands	$\underset{p}{\operatorname{argmax}} \int_{v^+} P(C(s), v^+ v)\ E(v^+, p, v)$	Paul *et al.* 2017
Paraphrasing	$\int_v \lvert P(s, v) - P(s', v) \rvert$	
Translation	$\underset{s' \in L'}{\operatorname{argmin}} \int_v \lvert P(s, v) - P(s', v) \rvert$	
Common sense reasoning	$\underset{s \in L}{\operatorname{argmax}} \int_v P(s_q, v)\ P(\mathrm{Q}(s, s_q), v)$	
Planning	$\underset{s \in L}{\operatorname{argmax}} \int_v P(s, v_0 : v : v_n)$	
…		

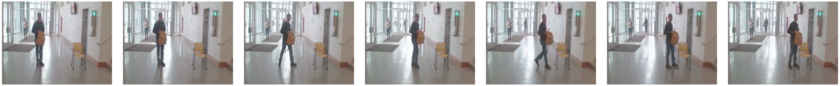
5g of brain

5g of brain

Planning

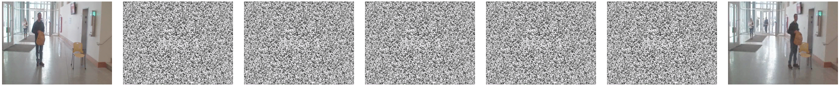
Planning

Planning



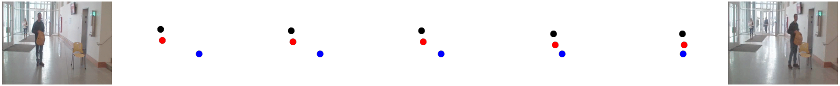
Danny carried the backpack to the chair.

Planning



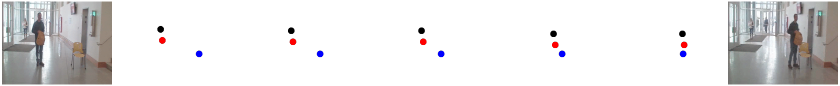
Danny carried the backpack to the chair.

Planning



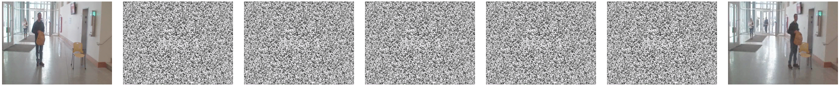
Danny carried the backpack to the chair.

Planning



Danny carried the backpack to the chair.

Planning



Danny carried the backpack to the chair.

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \text{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\operatorname*{argmax}_{\theta} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015
Follow commands	$\operatorname*{argmax}_{p} \int_{v^+} P(C(s), v^+ v)\ E(v^+, p, v)$	Paul *et al.* 2017
Paraphrasing	$\int_{v} \lvert P(s, v) - P(s', v) \rvert$	
Translation	$\operatorname*{argmin}_{s' \in L'} \int_{v} \lvert P(s, v) - P(s', v) \rvert$	
Common sense reasoning	$\operatorname*{argmax}_{s \in L} \int_{v} P(s_q, v)\ P(\mathrm{Q}(s, s_q), v)$	
Planning	$\operatorname*{argmax}_{s \in L} \int_{v} P(s, v_0 : v : v_n)$	

…

The long road ahead . . .

The long road ahead . . .

Coherent stories

The long road ahead . . .

Coherent stories

3D

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Physics

Physics

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Pooh

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

The long road ahead . . .

Coherent stories

3D

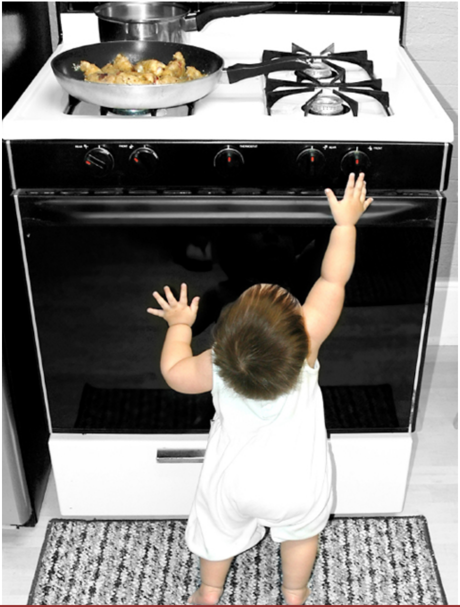
Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

A person standing in front of a stove



Theory of mind

Theory of mind

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

Social understanding

Social understanding

Social understanding

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

Social understanding

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

Social understanding

Modification

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

Social understanding

Modification

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

Social understanding

Modification

The vast majority of verbs: absolve, admire, anger, approve, bark, bend, chase, cheat, complete, concede, discover, fire, follow, fumble, hurry, race, recruit, reject, scratch, steal, taste, want, etc.

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

Social understanding

Modification

The vast majority of verbs: absolve, admire, anger, approve, bark, bend, chase, cheat, complete, concede, discover, fire, follow, fumble, hurry, race, recruit, reject, scratch, steal, taste, want, etc.

Metaphoric extension

The long road ahead . . .

Coherent stories

3D

Physics: Forces & contact relations

Segmentation

Parts and low-level features

Theory of mind

Social understanding

Modification

The vast majority of verbs: absolve, admire, anger, approve, bark, bend, chase, cheat, complete, concede, discover, fire, follow, fumble, hurry, race, recruit, reject, scratch, steal, taste, want, etc.

Metaphoric extension

etc.

Thanks to many great collaborators

Yevgeni Berzak, Candace Ross, Yen-Ling Kuo, Jonathan Malmaud
Daniel Harari, Battushig Myanganbayar, David Mayo, Nazar Ilamanov
Boris Katz, Shimon Ullman, Josh Tenenbaum

Siddharth Narayanaswamy, Jeffrey Siskind
Victor Carbarera, Santiago Perez
Sven Dickinson, Song Wang
Daniel Barrett, Haonan Yu
Maria Ryskina, Sergey Voronov

Recognition	$P(\text{sentence}, \text{video})$	Narayanaswamy *et al.* 2014
Retrieval	$\operatorname*{argmax}_{v \in V} P(s, v)$	Barret *et al.* 2016
Generation	$\operatorname*{argmax}_{s \in L} P(s, v)$	Yu *et al.* 2015, Narayanaswamy *et al.* 2
Question answering	$\operatorname*{argmax}_{s \in L} P(\mathrm{Q}(s, s_q), v)$	Barbu *et al.* in prep.
Disambiguation	$\operatorname*{argmax}_{i \in \text{parser}(s)} P(i, v)$	Berzak *et al.* 2015
Language acquisition	$\operatorname*{argmax}_{\theta} \prod_{s,v} P(s(\theta), v)$	Yu *et al.* 2015
Follow commands	$\operatorname*{argmax}_{p} \int_{v^+} P(C(s), v^+ v)\, E(v^+, p, v)$	Paul *et al.* 2017
Paraphrasing	$\int_v \lvert P(s, v) - P(s', v) \rvert$	
Translation	$\operatorname*{argmin}_{s' \in L'} \int_v \lvert P(s, v) - P(s', v) \rvert$	
Common sense reasoning	$\operatorname*{argmax}_{s \in L} \int_v P(s_q, v) \;\; P(\mathrm{Q}(s, s_q), v)$	
Planning	$\operatorname*{argmax}_{s \in L} \int_v P(s, v_0 : v : v_n)$	

…