CMPT 983

Grounded Natural Language Understanding

March 31, 2022

Speaker listener models

Today

- Bayesian models for color

- Rational Speech Acts (RSA)

Colors

Color test

- What color is this?

Color test

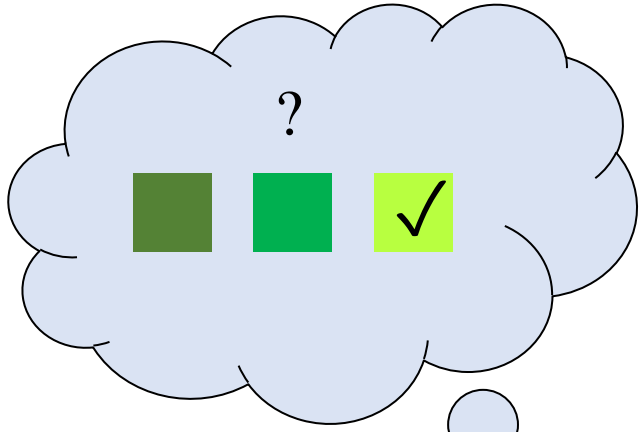
- What color is this?


Effective communications

- What you say depend on context and what the listener knows.
- Want to select words that are informative, clear and unambiguous.

Yellowish Green
Lime Green
Light Green

speaker



listener

Gricean Maxims

Guidelines for cooperative, effective communication

- Maxim of quantity: Give as much information as need, and no more
- Maxim of quality: Provide truthful information, supported by evidence
- Maxim of relation: Be relevant, say things pertinent to discussion
- Maxim of manner: Be clear, brief and orderly, avoid obscurity and ambiguity

To communicate clearly, we must have a shared convention of mapping of symbols to meanings.

Grounding color

Is there a true mapping of words to a single meaning?

- Given the same word, will two listeners have the same interpretation?

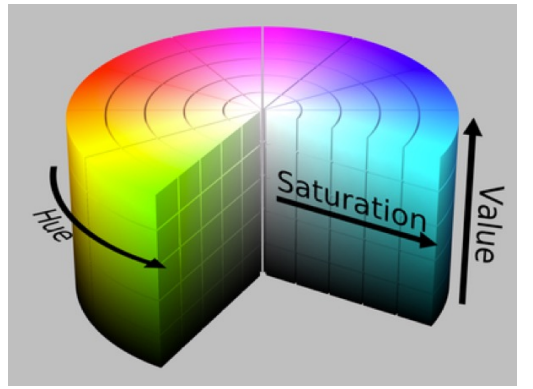
Green

- Given the same stimuli, will two speakers choose to use the same word?

(XKCD color survey, Randall Munroe, https://blog.xkcd.com/2010/05/03/color-survey-results/)

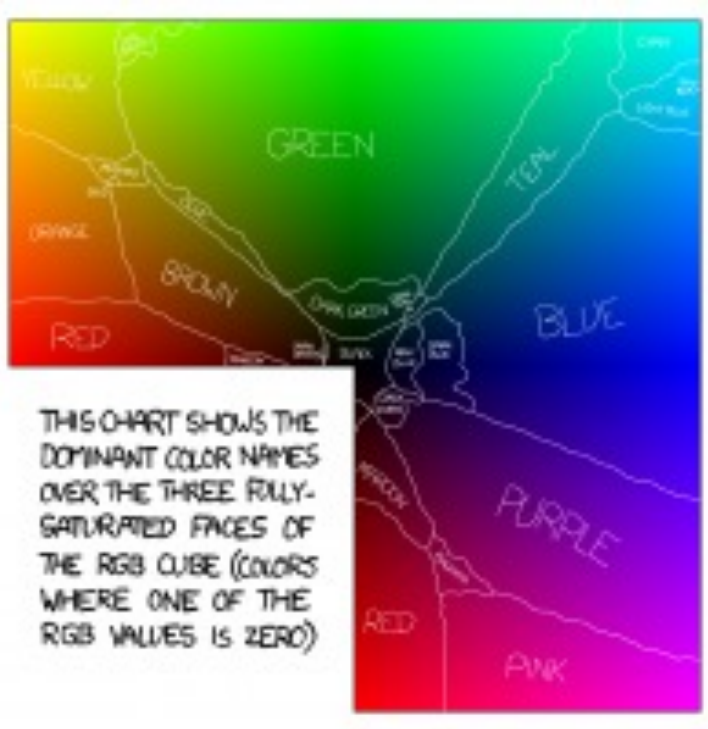
Grounding color

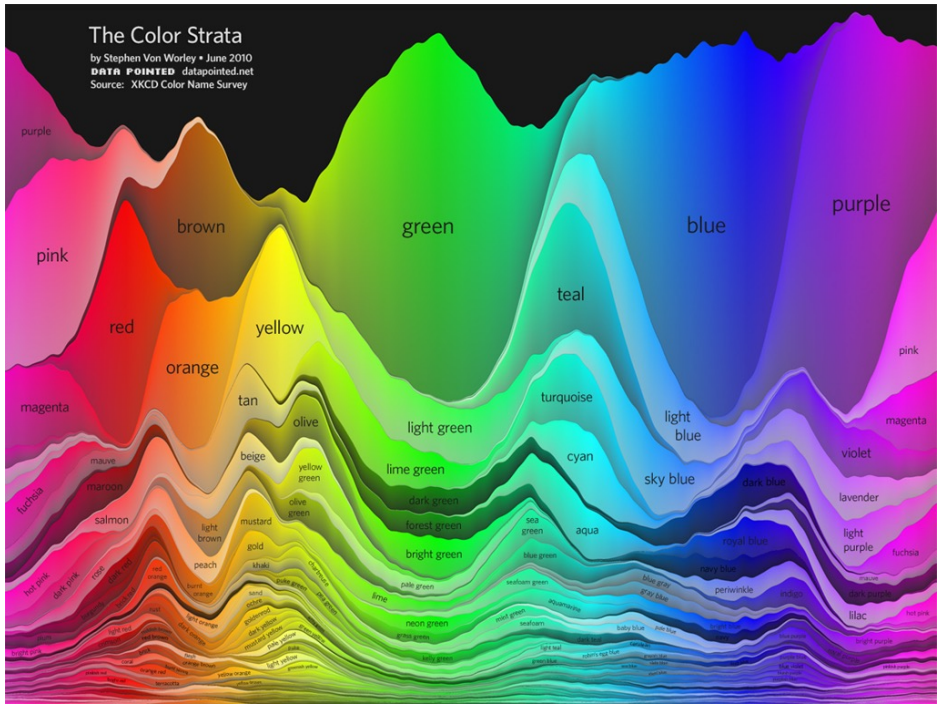
HSV color space



XKCD color survey

- Solicited names >5M random hues
- Got ~2.1M data points from >200K participants, with 829 distinct color names





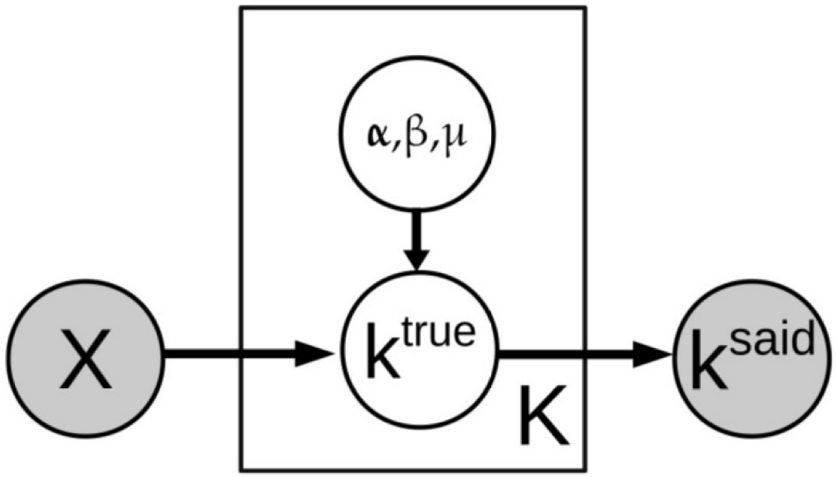
Let's use a probabilistic model!

(XKCD color survey, Randall Munroe, https://blog.xkcd.com/2010/05/03/color-survey-results/)

Grounding color

Bayesian model for grounded color semantics

- Model variation in meaning of words
- Given observed HSV color (X) and labels (k^{said}), how to learn a model of how to name colors?
- Speaker model: $P(k^{said} \mid X)$

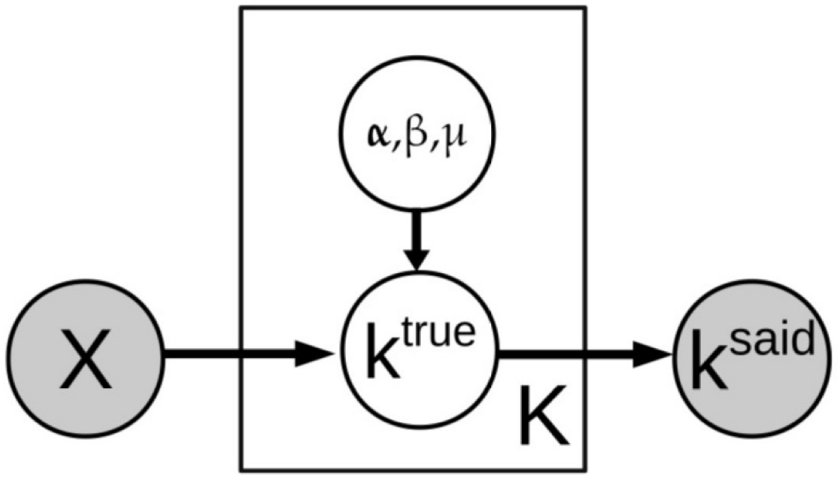


(A Bayesian Model of Grounded Color Semantics, McMahan and Stone, TACL 2015)

Grounding color

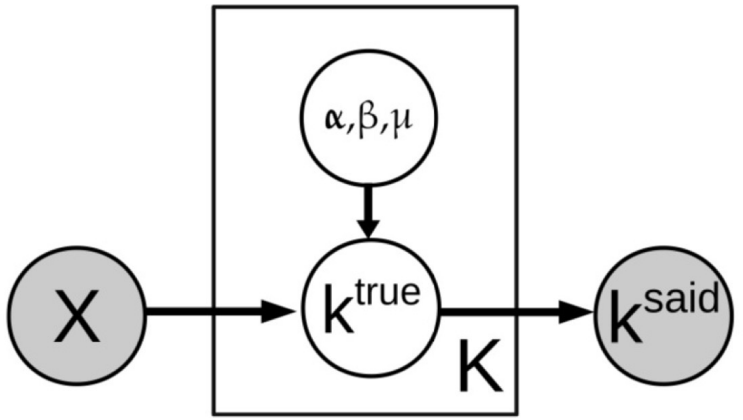
Bayesian model for grounded color semantics

- Model variation in meaning of words
- Given observed HSV color (X) and labels (k^{said}), how to learn a model of how to name colors?
- Speaker model: $P(k^{said} \mid X)$



(A Bayesian Model of Grounded Color Semantics, McMahan and Stone, TACL 2015)

Grounding color



Bayesian model for grounded color semantics

- Model variation in meaning of words
- Model probability distribution of color being called a given name

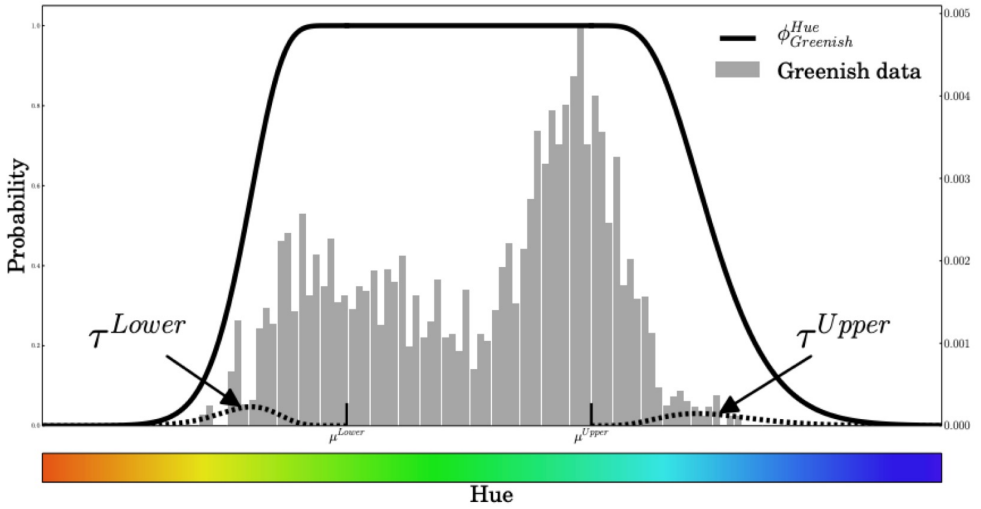
Model color channel (HSV) referred to by a color name *k* as a noisy box with a lower and upper threshold

$$\tau_k^{Lower,d} \sim \mu_k^{Lower,d} - \Gamma(\alpha_k^{Lower,d}, \beta_k^{Lower,d})$$
$$\tau_k^{Upper,d} \sim \mu_k^{Upper,d} + \Gamma(\alpha_k^{Upper,d}, \beta_k^{Upper,d})$$

Thresholds follow a gamma distribution from the mean for each dimension $d \in \{H, S, V\}$

Parameters estimated to maximize the log-likelihood of the Munroe color data

Greenish



(A Bayesian Model of Grounded Color Semantics, McMahan and Stone, TACL 2015)

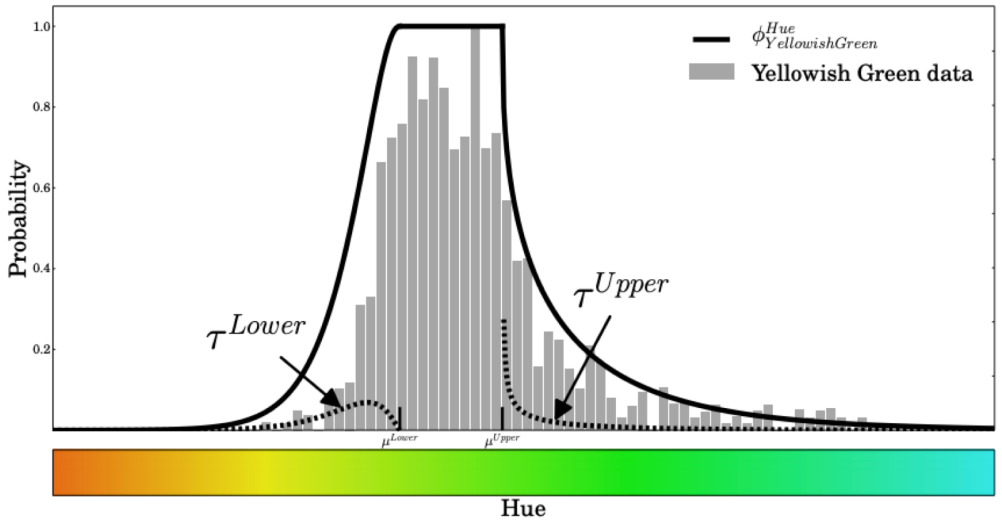
Grounding color

Lexicon of Uncertain Color Standards (LUX)
semantic representations of 827 English color labels

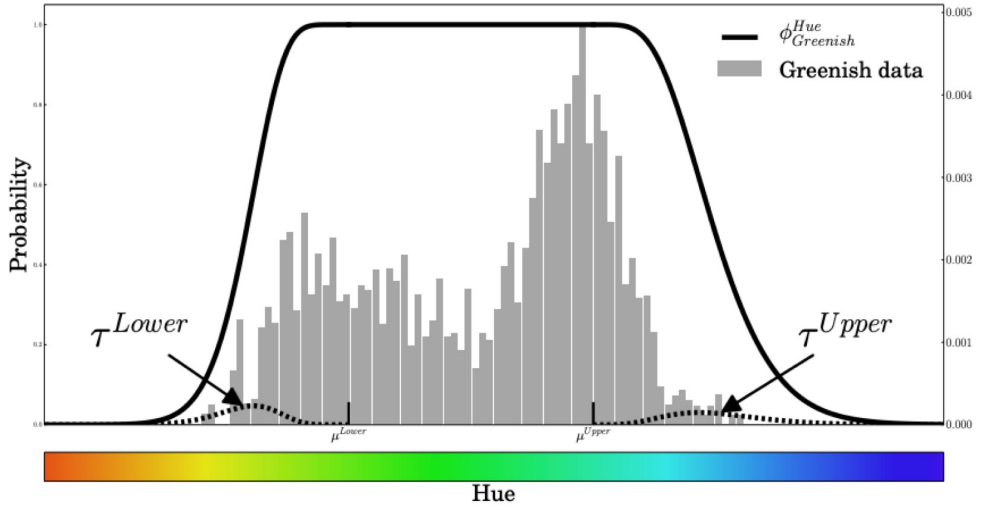
Bayesian model for grounded color semantics

- Model variation in meaning of words
- Probability distribution of denotation for each word

Yellowish Green



Greenish



(A Bayesian Model of Grounded Color Semantics, McMahan and Stone, TACL 2015)

Color test

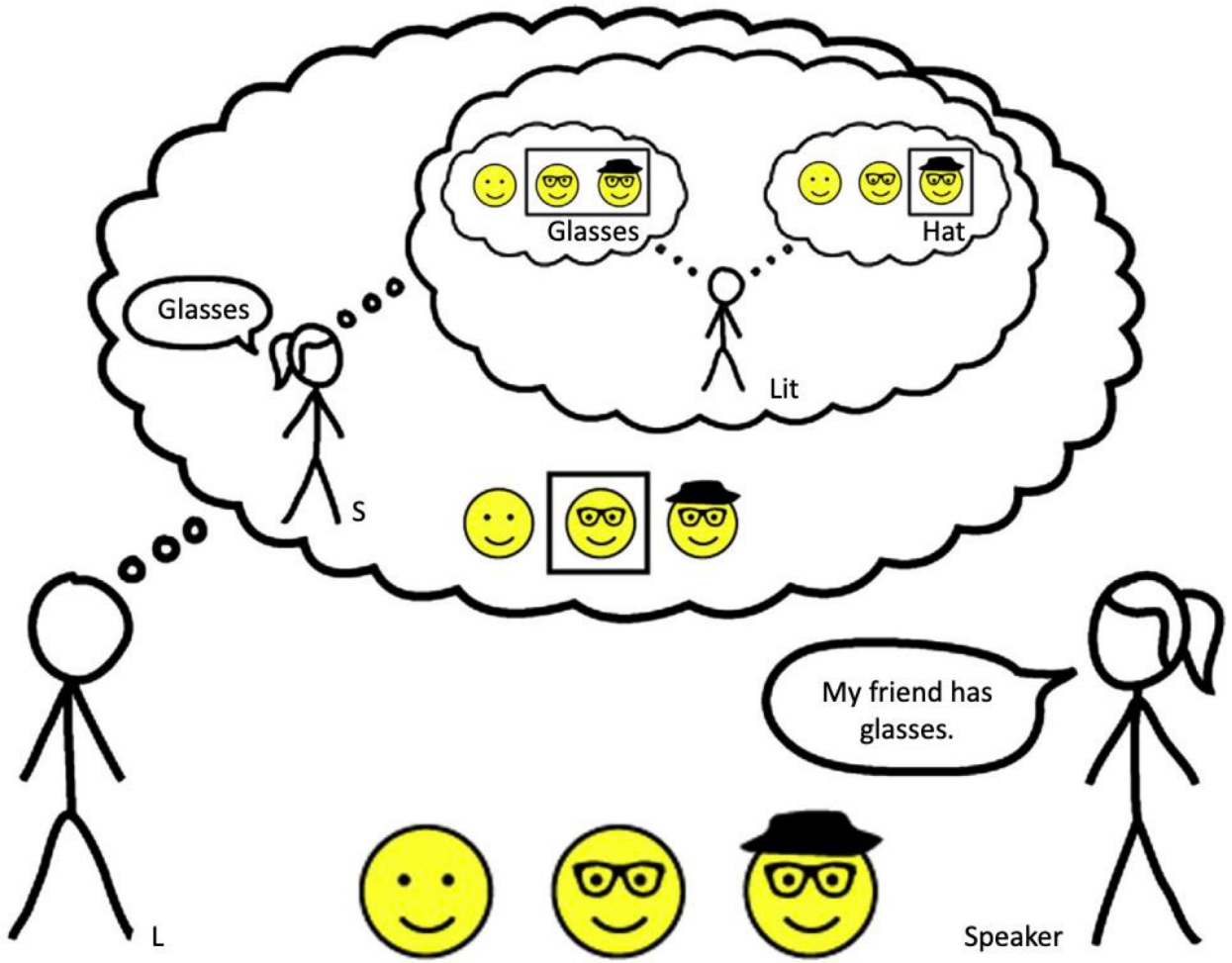
- What color is this?


What words would a speaker select to

- indicate each of these colors?
- so that the listener can pick out the correct color given the triplet?

Rational Speech Acts Framework

Probablistic Bayesian view



[Pragmatic Language Interpretation as Probabilistic Inference, Goodman and Frank 2016, http://langcog.stanford.edu/papers_new/goodman-2016-tics.pdf]

Reflex

Literal speaker and listeners

- Don't think about the other party
- Straightforward interpretation
- A bit of notation
  - u: utterance, t: world state,
  - M(u,t): meaning function connecting utterance u to world state t

    M(u,t) = 1 if u can be used to describe t, 0 otherwise

$M(u,t)$

u \ t		
blue	1	1
cyan	1	0

Assume uniform priors

$$S_0(u|t,M) \propto M(u,t)P(u)$$

$$L_0(t|u,M) \propto M(u,t)P(t)$$

speaker

u \ t		
blue	1/2	1
cyan	1/2	0

$S_0(u|t,M)$

u \ t		
blue	1/2	1/2
cyan	1	0

$L_0(t|u,M)$

listener

Example from *Understanding the Rational Speech Act model [Monroe et al, CogSci 2018]*

Pragmatic listener and speaker

- Pragmatics: how context contributes to meaning
  - any non-local meaning phenomena

    "Can you pass the salt?"

    "Is he 21?" "Yes, he's 25."

Literal version: "Can you pass the container with the salt in it?"

- Model mental state of the other party

Literal version: "Is he older than 21?"

Conversational implicatures

speaker

listener

Rational Pragmatic listener and speaker

$S_2(u|t, M)$

u \ t		
blue	1/4	1
cyan	3/4	0

$$S_2(u|t, M) \propto L_1(t|u, M)$$

$$S_0(u|t, M) \propto M(u, t)P(u)$$

$$L_1(t|u, M) \propto S_0(u|t, M)$$

speaker

u \ t		
blue	1/2	1
cyan	1/2	0

$S_0(u|t, M)$

u \ t		
blue	1/3	2/3
cyan	1	0

$L_1(t|u, M)$

listener

Example from *Understanding the Rational Speech Act model*

Pragmatic speaker and listener

$L_2(t|u, M)$

u \ t		
blue	1/4	3/4
cyan	1	0

$$L_2(t|u, M) \propto S_1(u|t, M)$$

$$S_1(u|t, M) \propto L_0(t|u, M)$$

u \ t		
blue	1/3	1
cyan	2/3	0

$S_1(u|t, M)$

$$L_0(t|u, M) \propto M(u, t)P(t)$$

u \ t		
blue	1/2	1/2
cyan	1	0

$L_0(t|u, M)$

speaker

listener

Example from *Understanding the Rational Speech Act model*

Converged speaker-listener model

After many iterations

u \ t		
blue	0	1
cyan	1	0

A more complex example

S_0

					
cyan	0.03	0	0	0	0.01
blue-green	0.02	0.01	0	0	0.01
blue-grey	0	0	0.01	0	0
blue-purple	0	0	0	0.01	0
bluish	0	0	0.01	0.01	0.02

S_n

					
cyan	0.21	0	0	0	0
blue-green	0.08	0.26	0	0	0.03
blue-grey	0	0	0.53	0	0
blue-purple	0	0	0	0.27	0
bluish	0	0	0	0	0.36

Example from *Understanding the Rational Speech Act model*

Moustache, Glasses, Hat example

$M(u,t)$

u \ t			
moustache	1	1	0
glasses	0	1	1
hat	0	0	1

$L_0(t|u,M)$

u \ t			
moustache	1/2	1/2	0
glasses	0	1/2	1/2
hat	0	0	1

$S_0(u|t,M)$

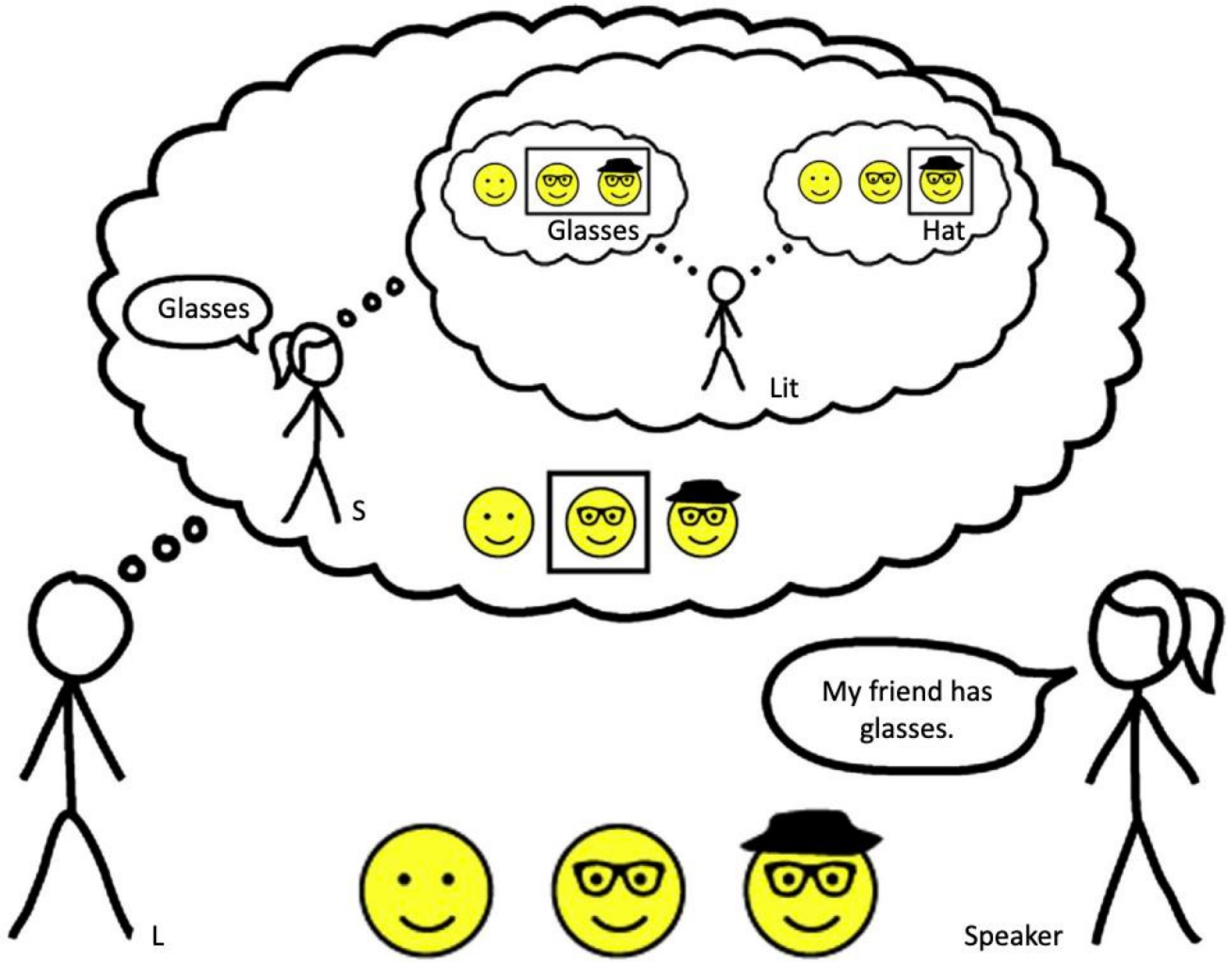
u \ t			
moustache	1	1/2	0
glasses	0	1/2	1/2
hat	0	0	1/2

Converged model

u \ t			
moustache	1	0	0
glasses	0	1	0
hat	0	0	1

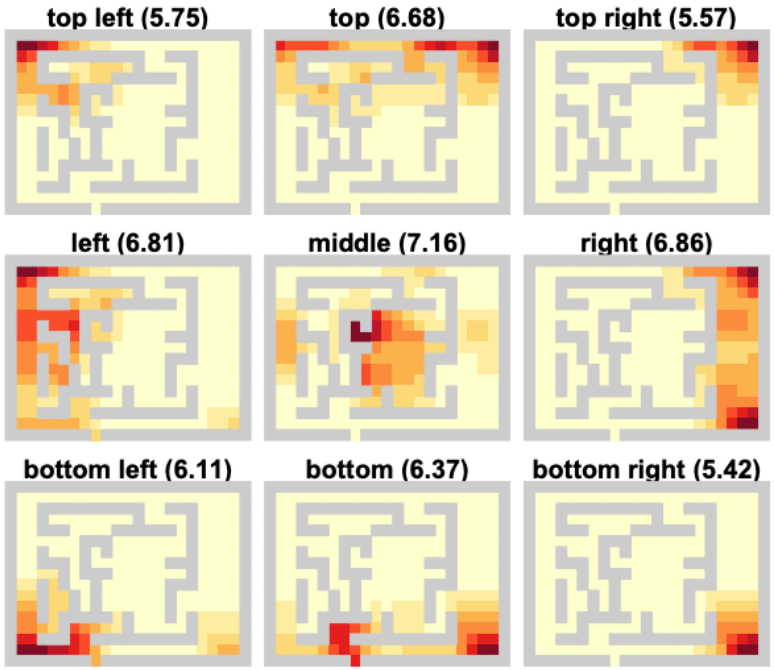
Example from *Implicatures and Nested Beliefs in Approximate Decentralized-POMDPs*, Vogel et al, ACL 2013

Do we need to keep recursing?



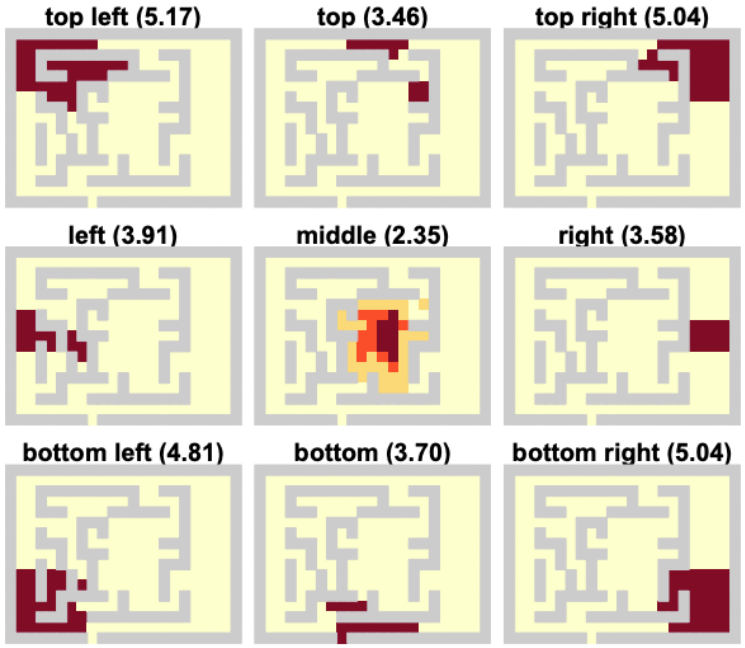
- Can be computationally expensive
- Let's consider basic level 1 speaker and listener models

Spatial references

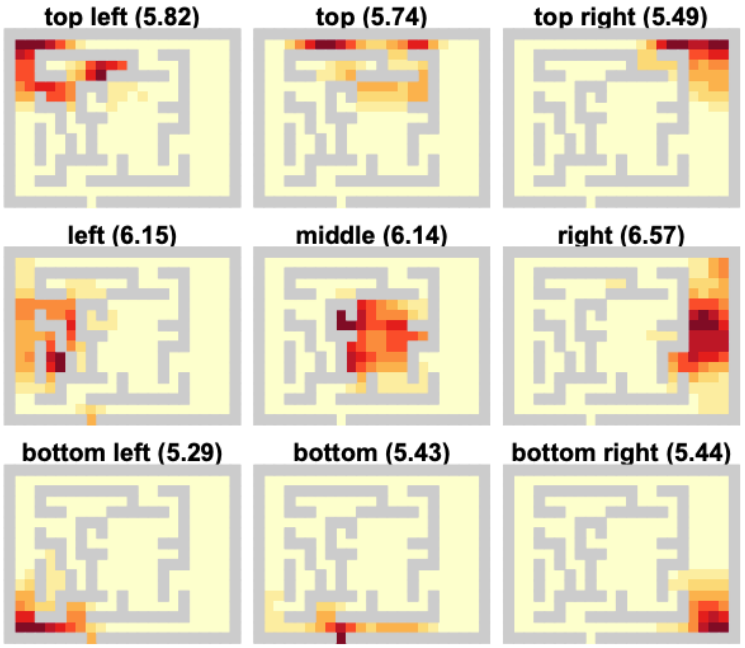


Literal listener (level 0)

Pragmatic listener (level 1)



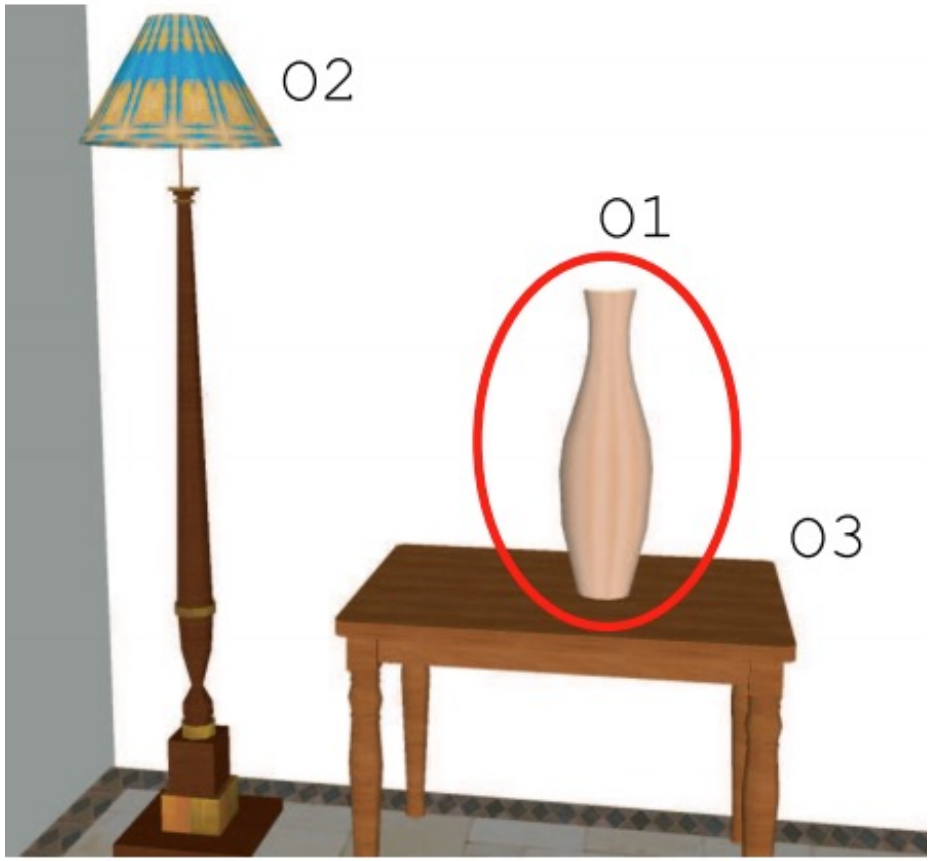
Human speakers



Example from *Implicatures and Nested Beliefs in Approximate Decentralized-POMDPs*, Vogel et al, ACL 2013

Speaker listener in applications (research papers)

Spatial relations



Consider only use spatial relations wrt to other objects to indicate (pick out) an object

- (i.e. do not say it is a vase or mention its color or other inherent properties)

How to indicate O1?

- Requires modeling listener
- "right of O2" is not sufficient to disambiguate the object

A Game-Theoretic Approach to Generating Spatial Descriptions, Golland et al, EMNLP 2010


Can you give me the orange book on top?

What to say?

the book?
the orange book?

Referring Expression
Generation

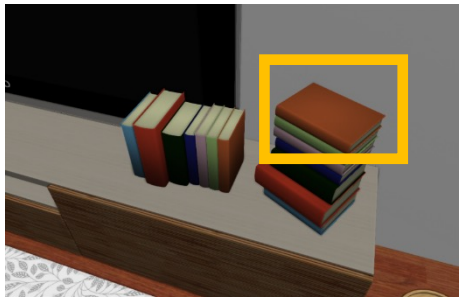
What did he mean?

Referring Expression
Comprehension

Need mental model of the other person

Referring expression generation

- Input: Image I with region R
- Output: Description S^*



orange book on top

$$S^* = \arg\max_S P(S|R, I)$$

L0 Speaker

Similar to standard image captioning task except input is a region in additional to the full image

- The full image / surrounding objects are used as context

Referring expression **comprehension**

- Input: Image I with description S

  Generate candidate regions C
- Output: Region R^*

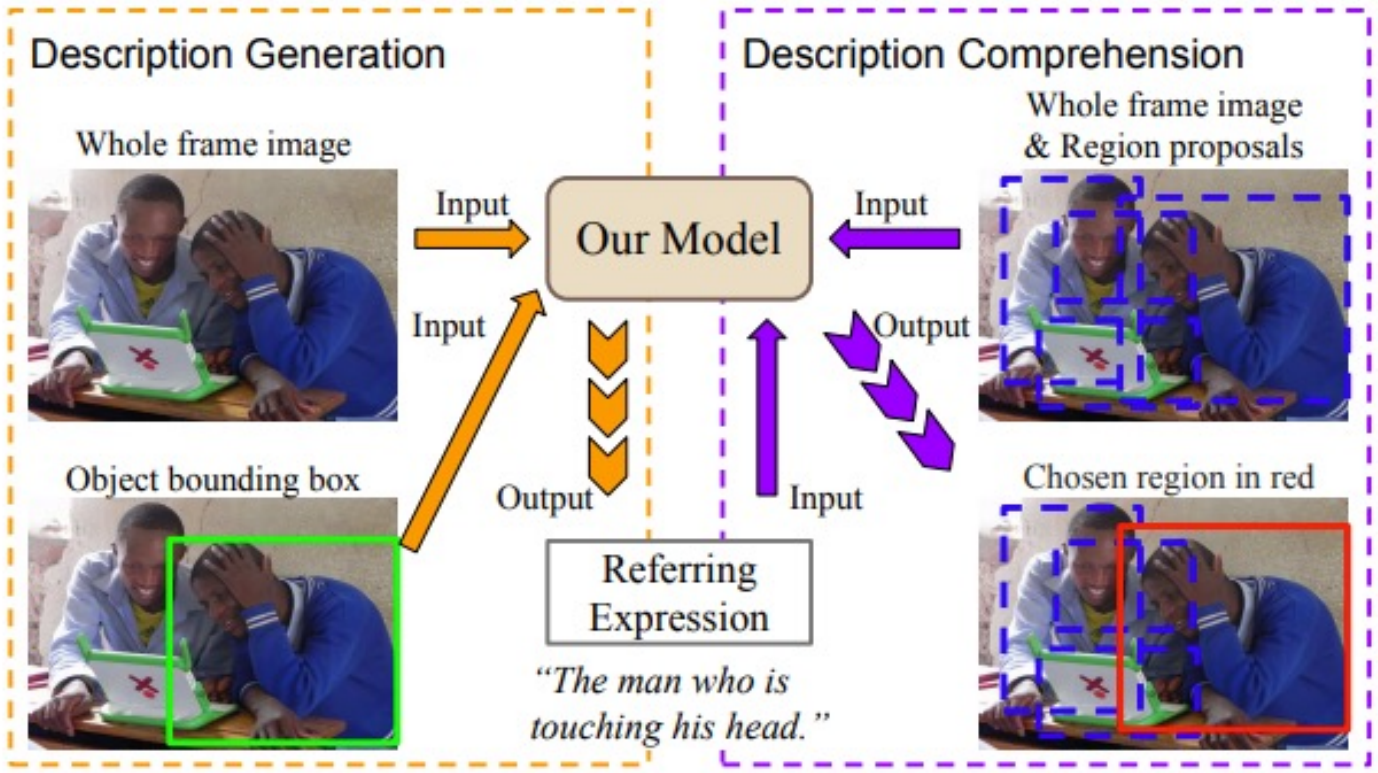
$$R^* = \arg\max_{R \in C} P(R|S, I)$$

Bayes Rule

$$P(R|S, I) = \frac{P(S|R, I)P(R|I)}{\sum_{R' \in C} P(S|R', I)P(R'|I)}$$

L1 Listener

Jointly modeling speakers and listeners for referring expressions



- Will training jointly result in more discriminative descriptions?

Generation and comprehension of unambiguous object descriptions, Mao et al, CVPR 2016

Jointly modeling speakers and listeners for referring expressions

Compare various training strategies

- Maximum Likelihood

$$J(\theta) = -\sum_{n=1}^{N} \log p(S_n|R_n, I_n, \theta)$$ L0 Speaker

- Maximum Mutual Information

$$J'(\theta) = -\sum_{n=1}^{N} \log p(R_n|S_n, I_n, \theta)$$ L1 Listener

Same as maximizing mutual information between speaker and listener (assuming $p(R)$ is uniform)

$$\text{MI}(S, R) = \log \frac{p(S, R)}{p(R)p(S)} = \log \frac{p(S|R)}{p(S)}$$

- Max-Margin MMI

$$J''(\theta) = -\sum_{n=1}^{N} \{\log p(S_n|R_n, I_n, \theta) - \max(0,\ M - \log p(S_n|R_n, I_n, \theta) + \log p(S_n|R'_n, I_n, \theta))\}$$

Evaluate using localization with GT vs generated description with IoU@0.5

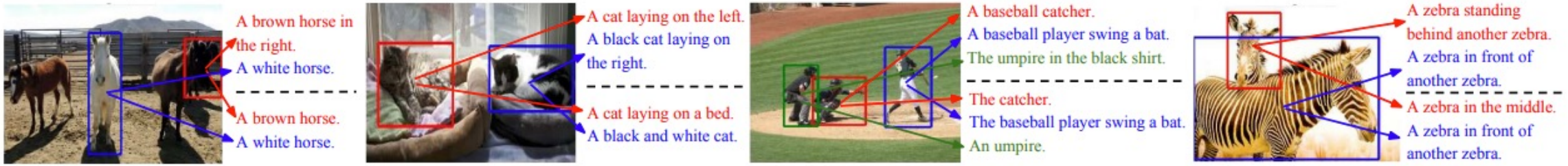
Proposals	GT		Multibox	
Descriptions	GEN	GT	GEN	GT
ML (baseline)	0.803	0.654	0.564	0.478
MMI-MM-easy-GT-neg	0.851	0.677	0.590	0.492
MMI-MM-hard-GT-neg	**0.857**	**0.699**	0.591	0.503
MMI-MM-multibox-neg	0.848	0.695	**0.604**	**0.511**
MMI-SoftMax	0.848	0.689	0.591	0.502

Generation and comprehension of unambiguous object descriptions, Mao et al, CVPR 2016

Jointly modeling speakers and listeners for referring expressions

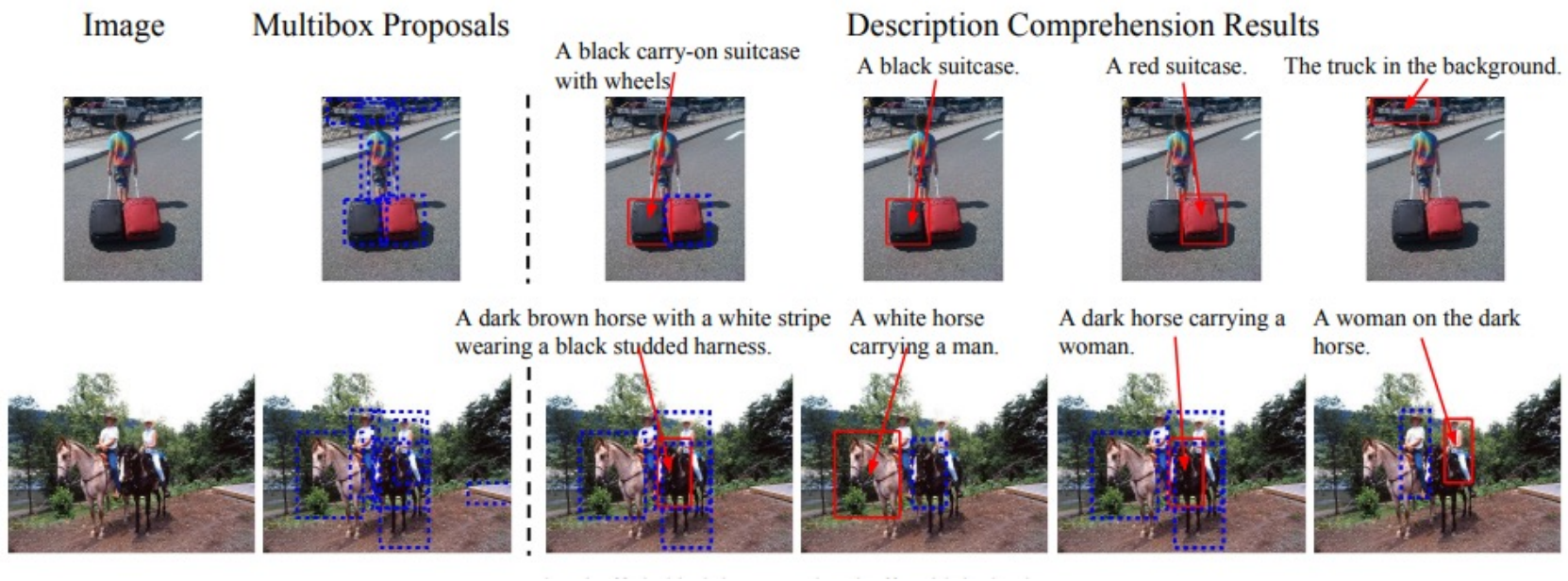
Example generated captions

Full model



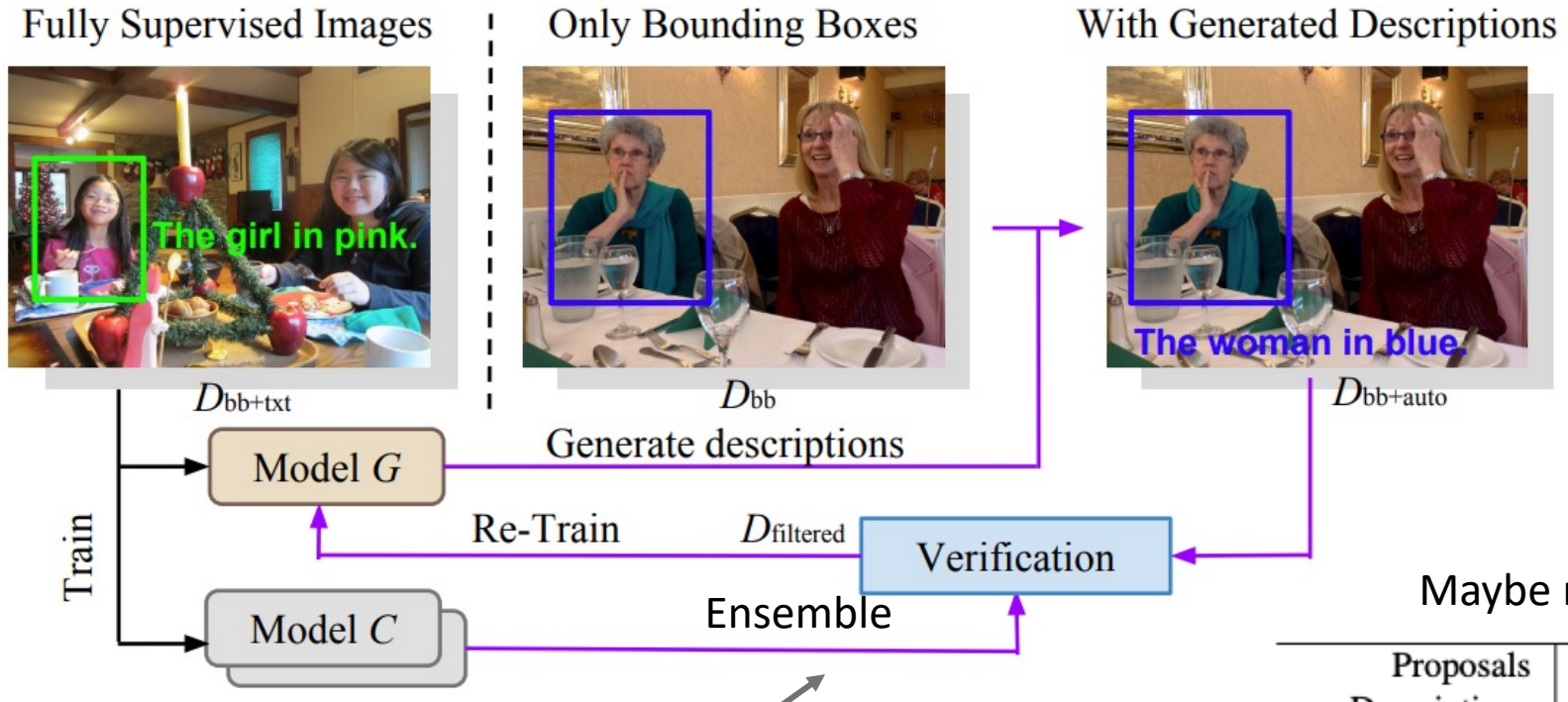
Baseline

Example localizations



Generation and comprehension of unambiguous object descriptions, Mao et al, CVPR 2016

Jointly modeling speakers and listeners for referring expressions



- Semi-supervised training
- Generated descriptions bounding boxes

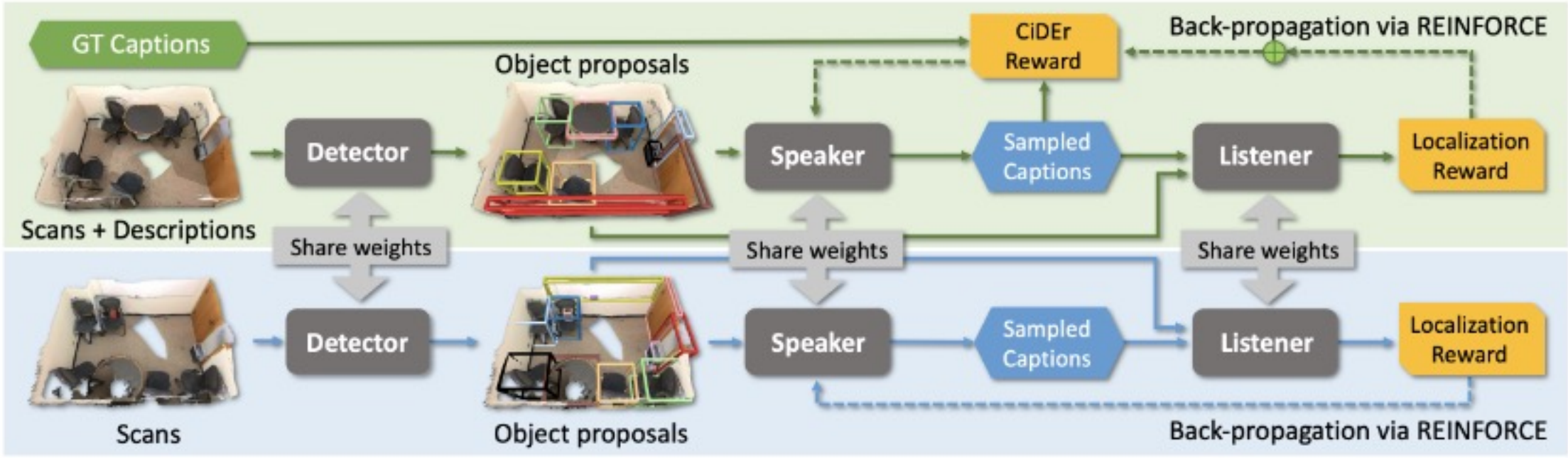
Maybe not help that much

Proposals	GT		multibox	
Descriptions	GEN	GT	GEN	GT
G-Ref				
D_{bb+txt}	0.791	0.561	**0.489**	0.417
$D_{bb+txt} \cup D_{bb}$	**0.793**	**0.577**	**0.489**	**0.424**
UNC-Ref				
D_{bb+txt}	0.826	0.655	0.588	0.483
$D_{bb+txt} \cup D_{bb}$	**0.833**	**0.660**	**0.591**	**0.486**

Generation and comprehension of unambiguous object descriptions, Mao et al, CVPR 2016

Visual grounding in 3D scans

- Unified framework that detects objects and describes (speaker) and discriminates (listener)
- Speaker can be used to generate captions for objects without annotated descriptions



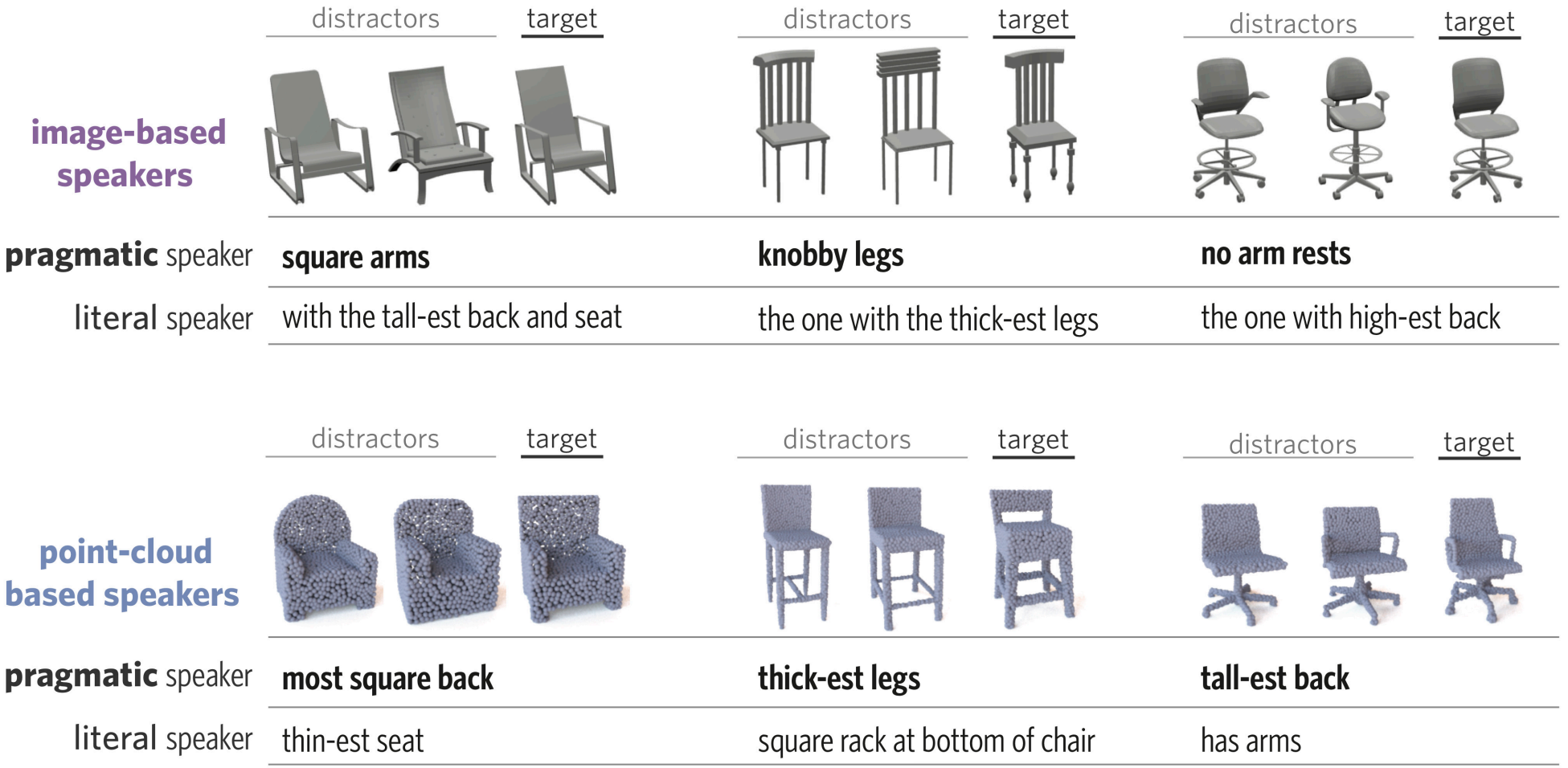
D^{3}Net: A Speaker-Listener Architecture for Semi-supervised Captioning and Visual Grounding in RGB-D Scans, Chen et al, 2021

Visual grounding in 3D scans

	Captioning Metrics (CiDER, BLEU-4, METEOR, ROUGE)				Localization
	C@0.5IoU	B-4@0.5IoU	M@0.5IoU	R@0.5IoU	mAP@0.5
Scan2Cap	39.08	23.32	21.97	44.78	36.13
Ours (MLE)	46.07	30.29	24.35	51.67	50.93
Ours (CiDEr)	60.93	34.36	25.12	52.26	51.44
Ours (CiDEr+loc.)	61.30	34.50	25.25	52.80	52.07
Ours (CiDEr+loc.+lobjcls.)	61.50	35.05	25.48	53.31	52.58
Ours (w/ 0.1x extra data)	61.91	35.03	25.38	53.25	52.64
Ours (w/ 0.5x extra data)	62.36	35.54	25.43	53.67	53.17
Ours (w/ 1x extra data)	**62.64**	**35.68**	**25.72**	**53.90**	**53.97**

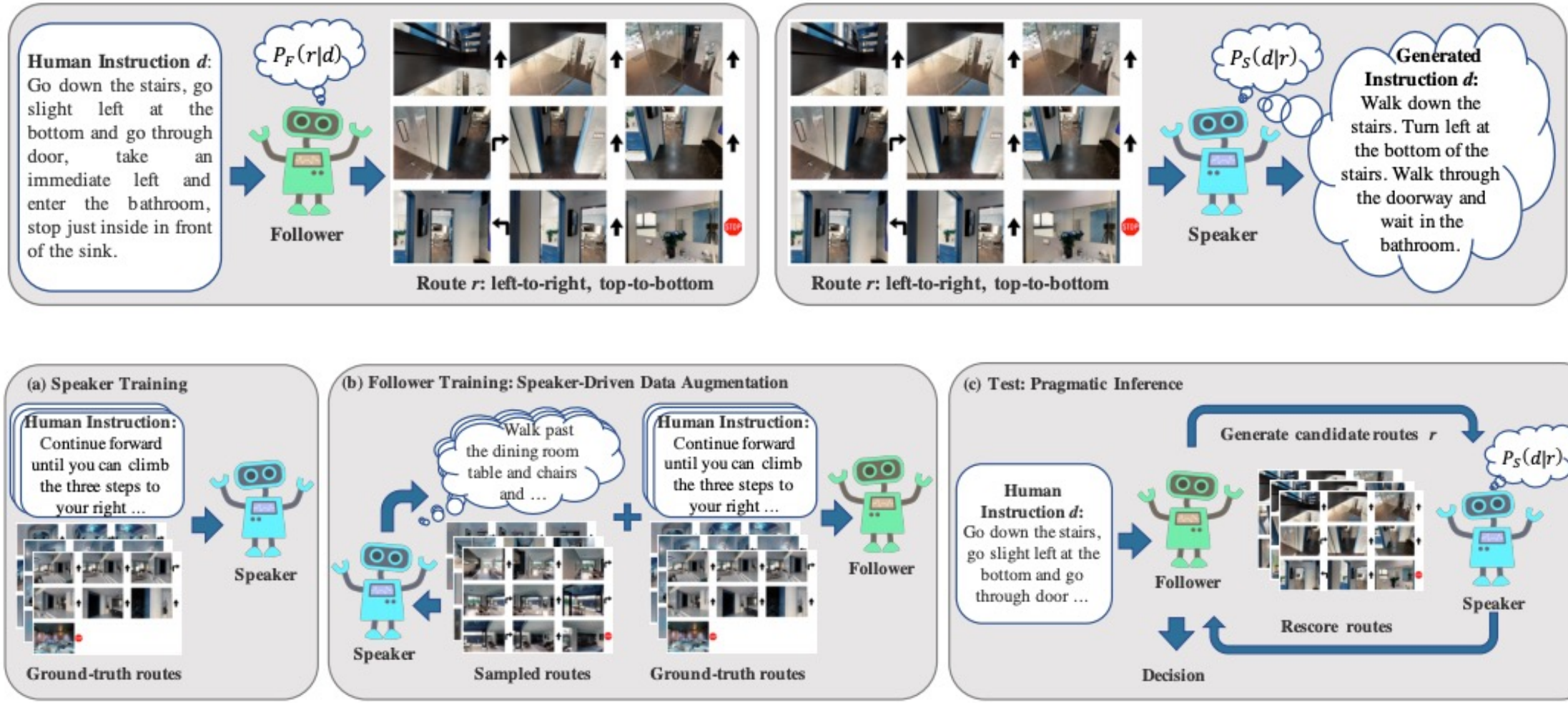
D^{3}Net: A Speaker-Listener Architecture for Semi-supervised Captioning and Visual Grounding in RGB-D Scans, Chen et al, 2021

ShapeGlot



ShapeGlot: Learning Language for Shape Differentiation, Achlioptas et al, ICCV 2019

Vision-language navigation



Speaker-Follower Models for Vision-and-Language Navigation, Fried et al, NIPS 2018

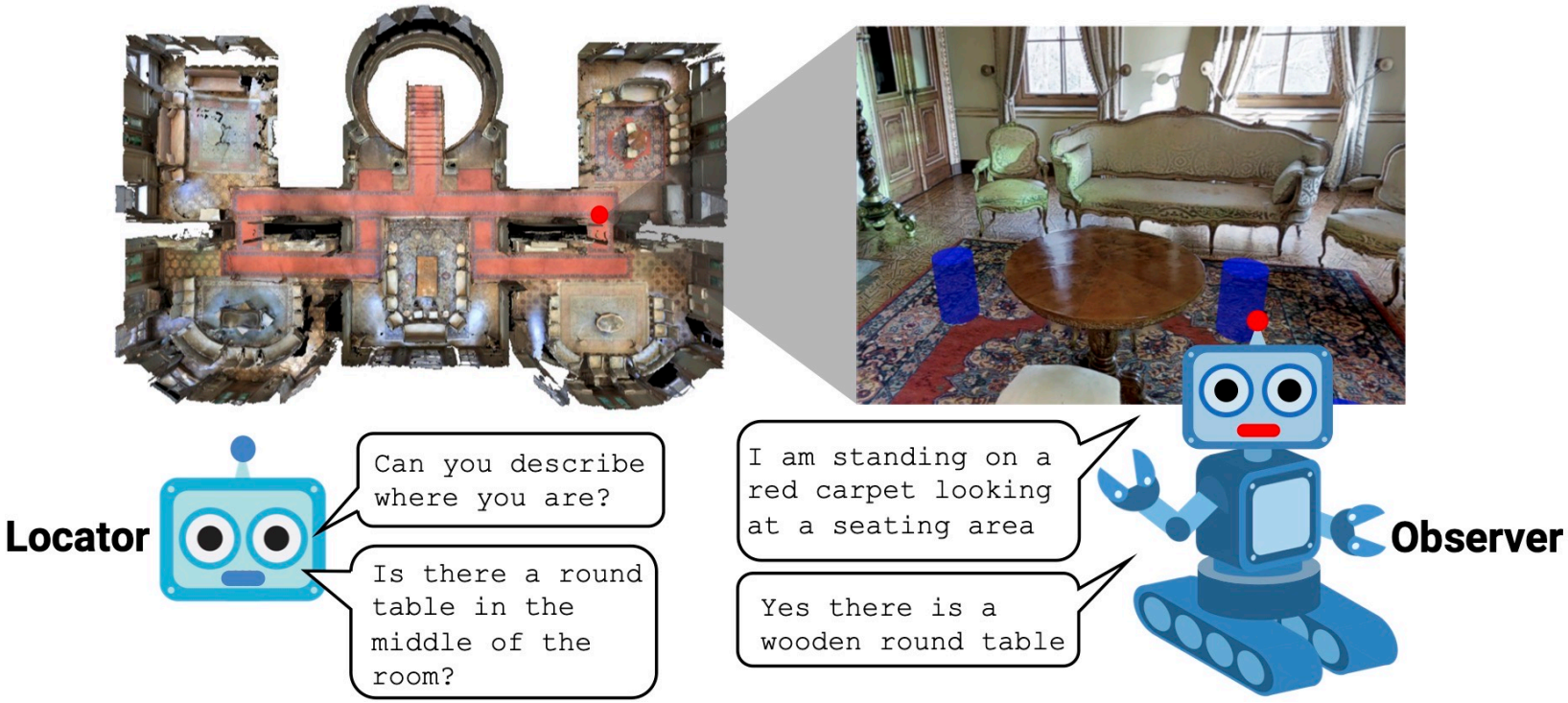
Data-augmentation for VLN

#	Data Augmentation	Pragmatic Inference	Panoramic Space	Validation-Seen NE ↓	Validation-Seen SR ↑	Validation-Seen OSR ↑	Validation-Unseen NE ↓	Validation-Unseen SR ↑	Validation-Unseen OSR ↑
1				6.08	40.3	51.6	7.90	19.9	26.1
2	✓			5.05	46.8	59.9	7.30	24.6	33.2
3		✓		5.23	51.5	60.8	6.62	34.5	43.1
4			✓	4.86	52.1	63.3	7.07	31.2	41.3
5	✓	✓		4.28	57.2	63.9	5.75	39.3	47.0
6	✓		✓	3.36	66.4	73.8	6.62	35.5	45.0
7		✓	✓	3.88	63.3	71.0	5.24	49.5	63.4
8	✓	✓	✓	**3.08**	**70.1**	**78.3**	**4.83**	**54.6**	**65.2**

Speaker-Follower Models for Vision-and-Language Navigation, Fried et al, NIPS 2018

Cooperative task: localization

- Asymmetric knowledge and abilities
- Goal: Locater guess where the observer is



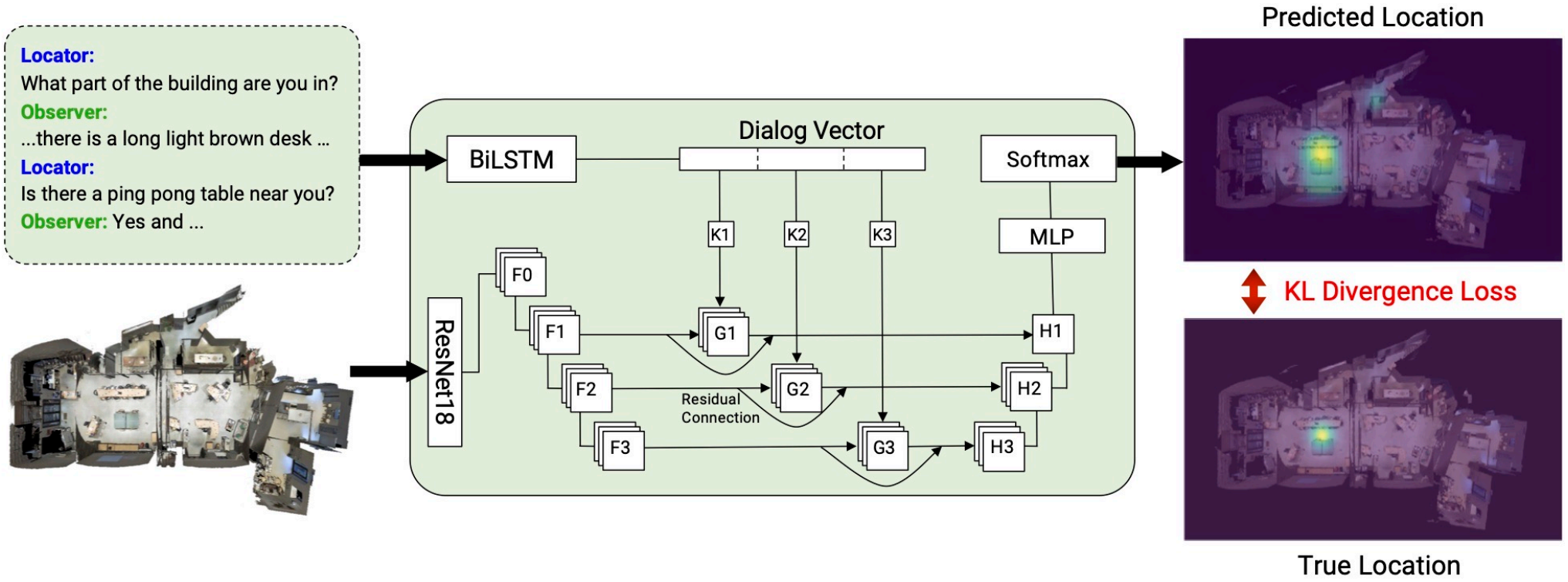
Locator: Has access to top down map, need to predict current position of the observer

Observer: Egocentric view, can move around

Where Are You? Localization from Embodied Dialog, Hahn et al, EMNLP 2020

LingUNet based architecture

- Convolutional encoder-decoder with language modulated skip-connections
  - Convolution weights are predicted from dialog encoding
  - Trained to minimize KL-divergence between GT and predicted locations



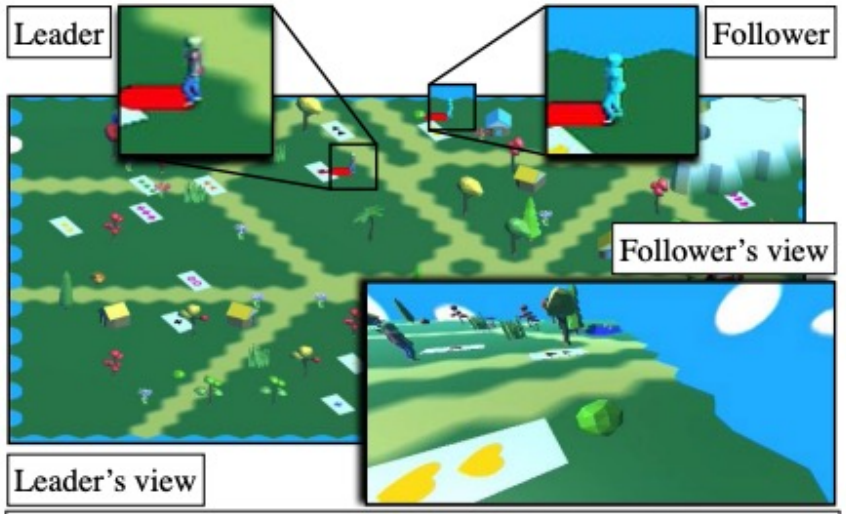
Where Are You? Localization from Embodied Dialog, Hahn et al, EMNLP 2020

Collaborative task: gather cards

- Asymmetric knowledge and abilities
- Goal: Leader + follower collect a set (3 cards)
  - Get a point (+ more turns) for every set (distinct shape, color, count)

Leader: Can see entire environments but has less steps per turn

Both: Can take actions (move, stop) and communicate (special action)

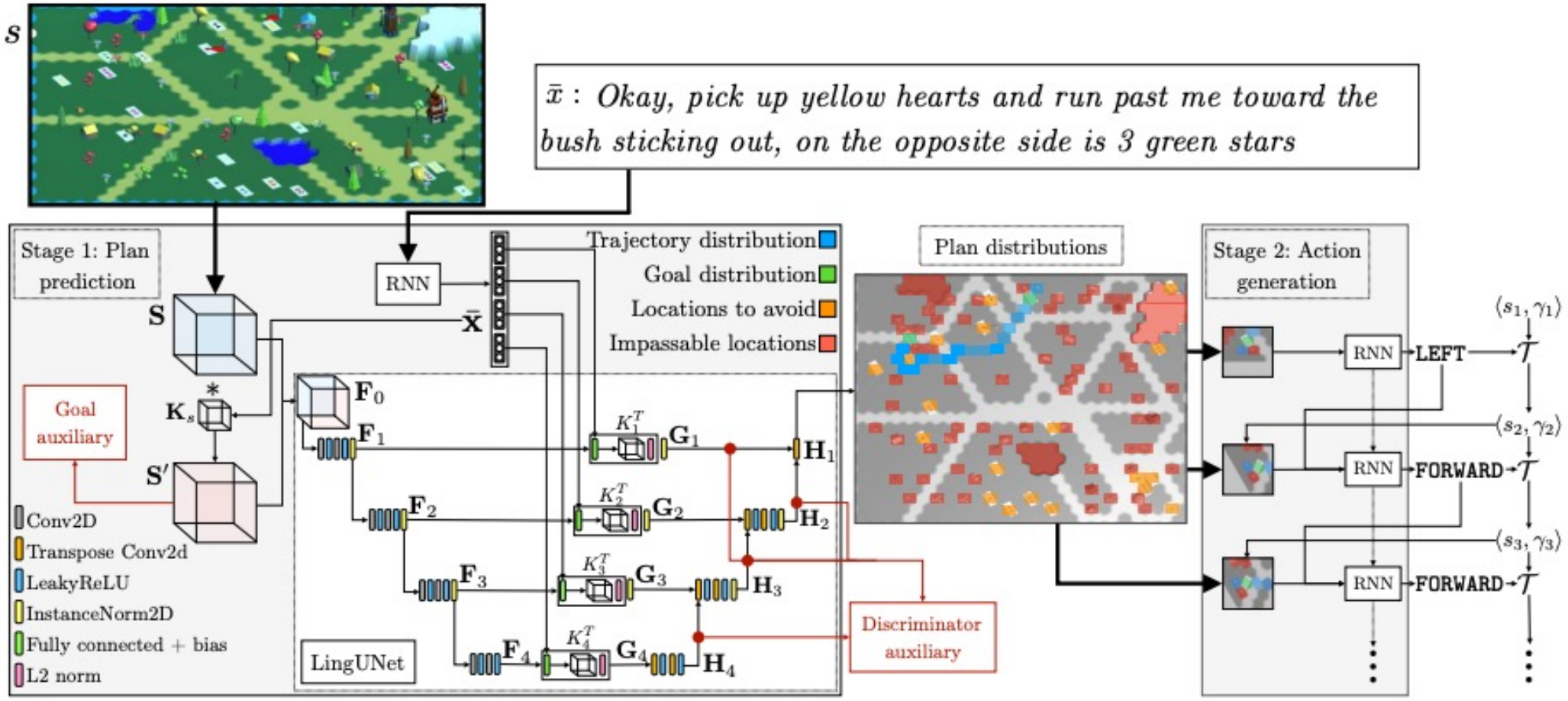


Follower: Egocentric view, 3 more steps per turn than leader.

Executing Instructions in Situated Collaborative Interactions, Suhr et al, EMNLP 2019

LingUNet based architecture

- Predict where to go next and generate action based on that



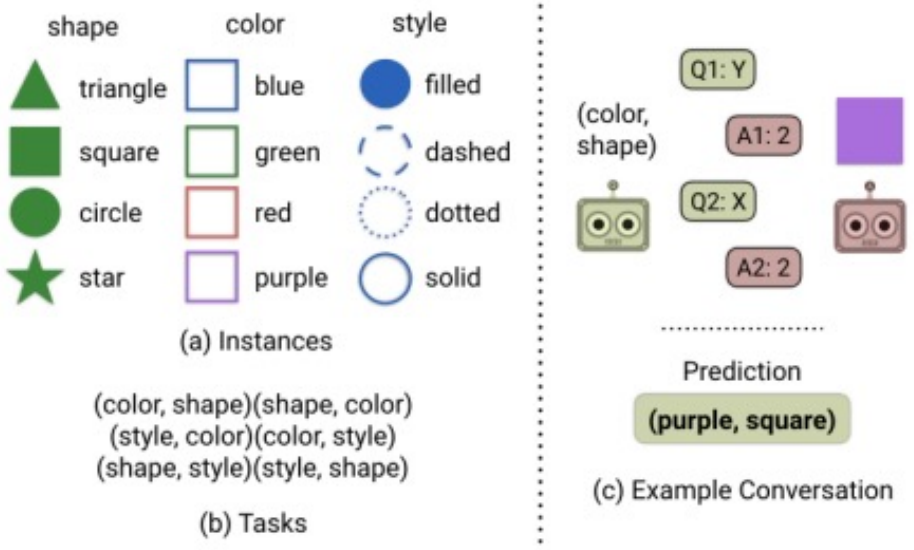
Executing Instructions in Situated Collaborative Interactions, Suhr et al, EMNLP 2019

Multi-agent communication

- Simulate speakers and listeners and see what happens
- Emergent communications!
- Collaborate task: two bots communicate about simple shapes
  - A-Bot: access to colored shape with a specific style
  - Q-Bot: knows specific task (attributes it needs to output) and ask A-Bot questions until it can guess the 2 attributes of interest

Goal: Guess ordered attributes
Communication: Some token from a vocabulary

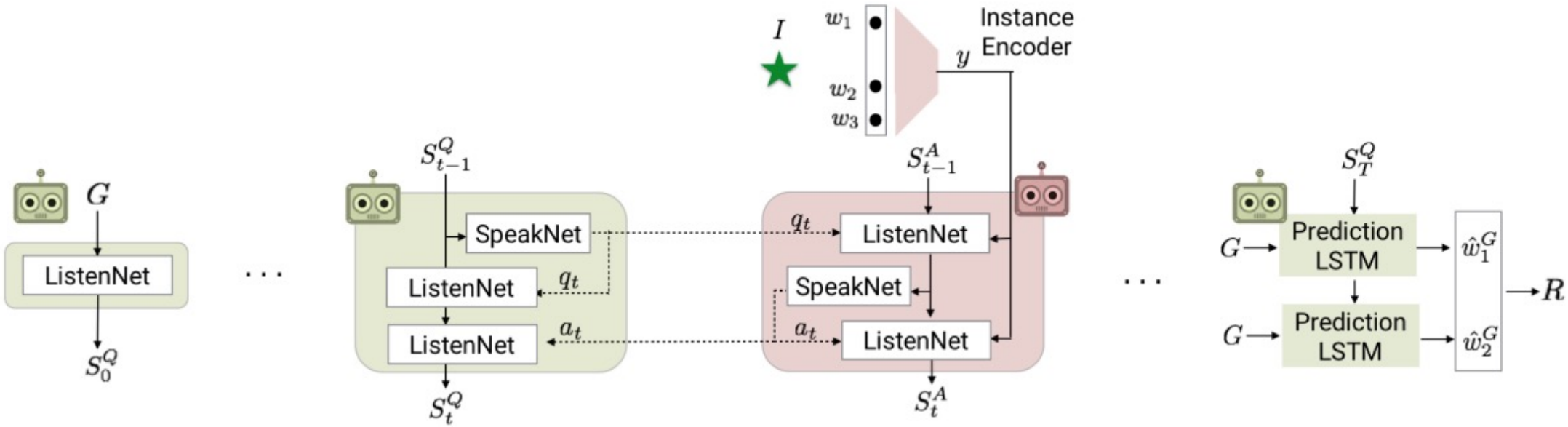
Total of 64 possible shape/color/style combinations



Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog, Kottur et al, EMNLP 2017

Multi-agent communication

- Each Bot has speaker and listener components
- Separate vocabularies for Q-Bot (green) and A-Bot (pink)
- Two rounds of communication
- Trained using REINFORCE (maximize Q-Bot accuracy)

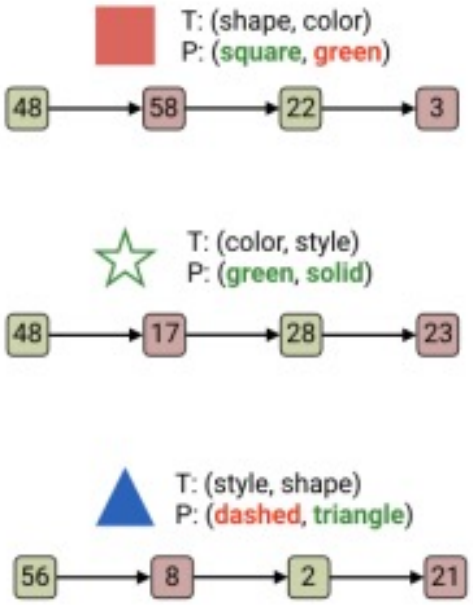


Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog, Kottur et al, EMNLP 2017

Multi-agent communication

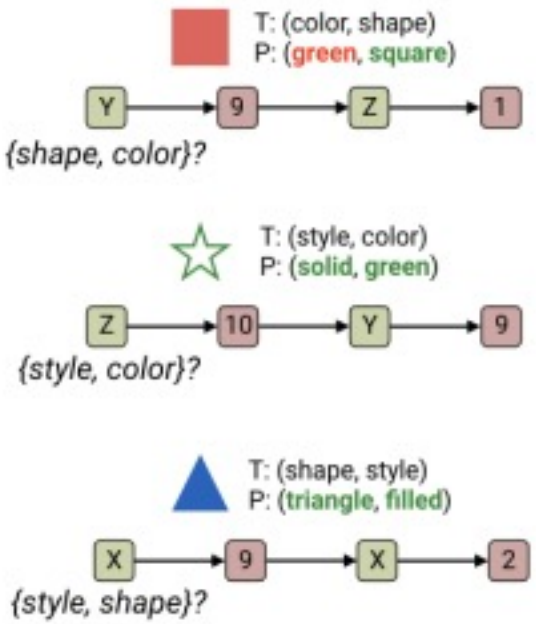
- Key Findings
  - Compositionality emerges only with restricted vocabulary

Large vocabulary ($|V_Q|$ = $|V_A|$ = 64)



A-Bot ignores Q-Bot, communicate using large "codebook of" instance to tokens, multiple token pairs can map to same instance

Small vocabulary ($|V_Q|$ = 3, $|V_A|$ = 12)

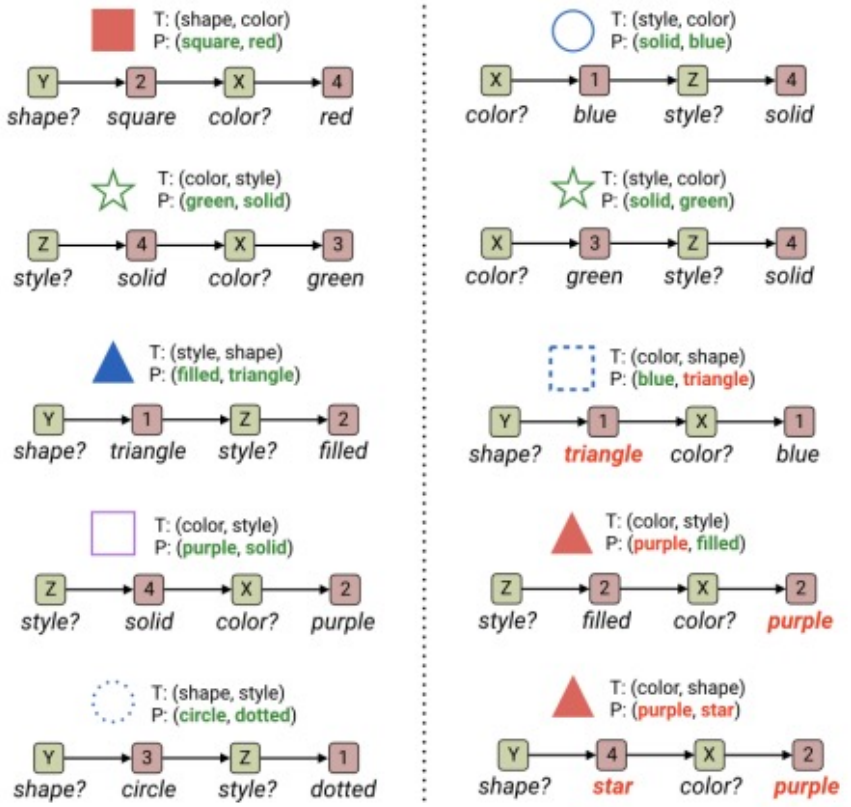


Q-Bot use first round to communicate task (ignoring order). A-Bot use set partitioning strategy to convey 16 options (meaning of tokens change across rounds).

Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog, Kottur et al, EMNLP 2017

Multi-agent communication

- Memory-less A-bot
  - Consistent and grounded language emerges



Small vocabulary ($|V_Q| = 3$, $|V_A| = 4$)

	Attributes			Task	q_1, q_2
V_A	*color* X	*shape* Y	*style* Z	*(color, shape)* *(shape, color)*	*Y, X*
1	*blue*	*triangle*	*dotted*	*(shape, style)* *(style, shape)*	*Y, Z*
2	*purple*	*square*	*filled*		
3	*green*	*circle*	*dashed*	*(color, style)*	*Z, X*
4	*red*	*start*	*solid*	*(style, color)*	*X, Z*

(a) A-BOT (b) Q-BOT

Natural Language Does Not Emerge 'Naturally' in Multi-Agent Dialog, Kottur et al, EMNLP 2017

Summary

- Speaker-listener
- RSA: Mental model of the other agent
- Full model computationally expensive and may not be necessary
- Simulate speakers and listener → emergent communications

Next time

- Paper presentations (4/4)
  - CLIPORT: What and Where Pathways for Robotic Manipulation (Michael)
  - The Emergence of the Shape Bias Results from Communicative Efficiency (Brian)

- Wednesday (4/6): Project presentations