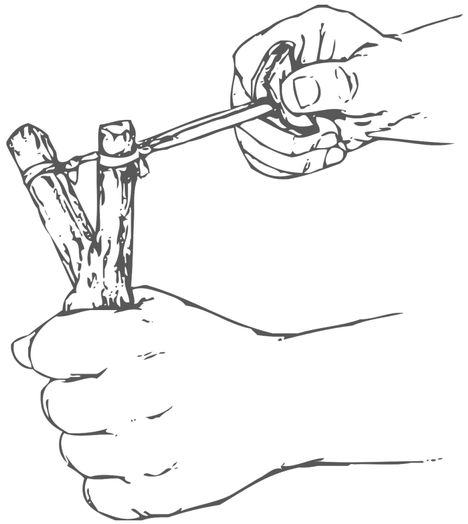
Neural·Pragmatic
Natural
Language
Generation

N·P
NLG

Learning goals



1. become oriented in the landscape of pragmatic neural NLG
2. understand different ways in which RSA(-like) ideas can be applied in NLG:
   a. during training
   b. during inference

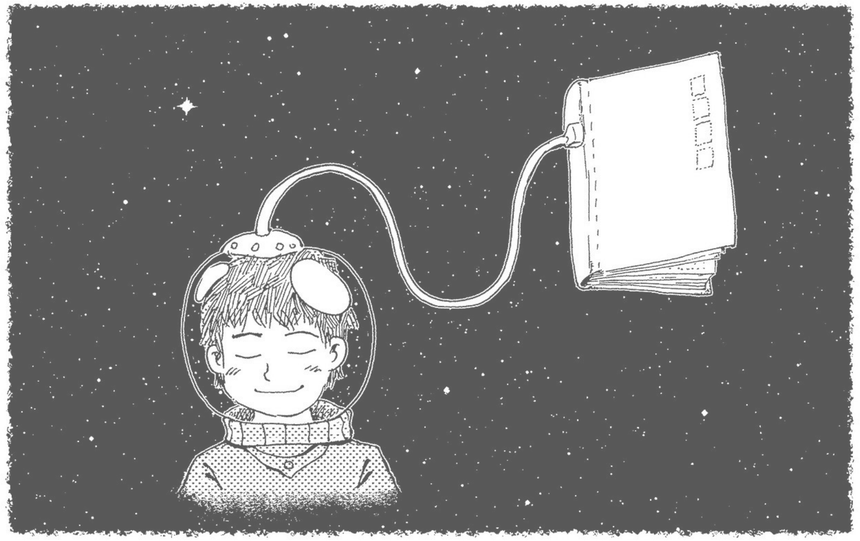
organizational remarks

Course projects

- work in groups (2-3 people are ideal)
  - single-person projects are okay but need motivation & permission
  - problems in the group discussed w/ lecturer before escalation
  - there will be one grade for the whole group
- outcome of the project
  - structured, documented, self-contained repository w/ all materials
  - highly accessible (reproducible, commented …) code
  - short research paper (PDF) explaining what was done, how this relates the to literature, why it was done and what was achieved or found
- content & scope
  - critical conceptual / mathematical work (even w/o any code) is welcome
  - typical project will aim to reproduce key results from a single paper
  - ambitious projects can shine by additionally:
    - extending or combining existing analyses
    - critically discussing existing analyses (in the light of the literature or project results)
    - conceptually motivated exploration of novel models, different data sets, other evaluation measures …

How to read a research paper

- identify key innovation / argument / point of the paper
  - how novel or important is this?
- track what you like and dislike
  - e.g., what's well explained, what's incomprehensible?
  - how can you incorporate what's good into your own repertoire?
  - how would *you* have done it differently?
- track what / how much you understand
  - what would I need in addition to understand more?
  - what don't I understand that I don't need to understand?
- take notes
  - organize and revisit your notes



RSA meets neural NLG

Pragmatic back-and-forth reasoning

speaker and listener reason about each other's behavior in a share context

speaker

$s_1, s_2, s_3, s_4, \ldots$

states

$P_S(u \mid s)$

$P_L(s \mid u)$

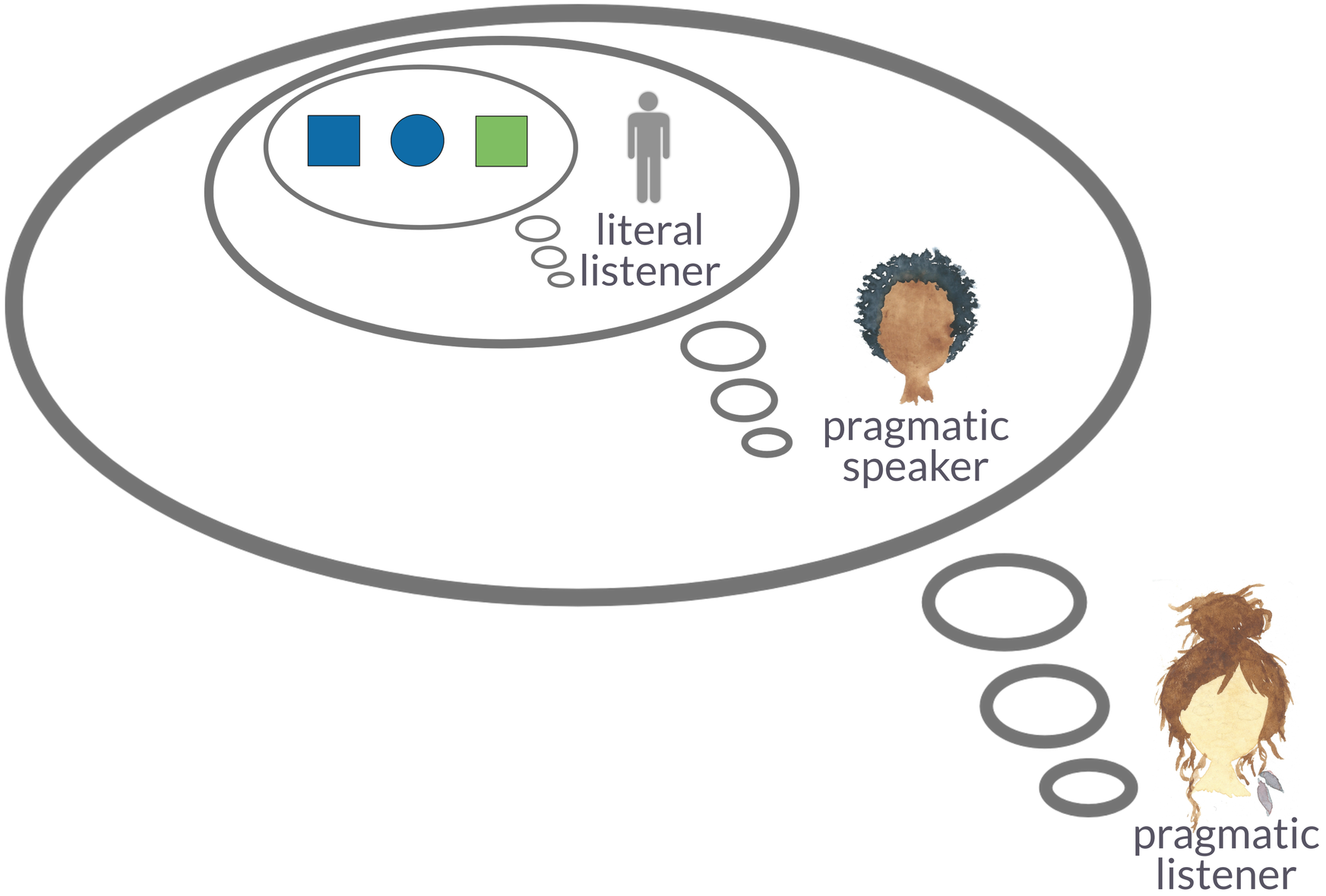
listener

$u_1, u_2, u_3, \ldots$

utterance

Grounding pragmatic reasoning

in a (dummy) literal listener



RSA-style
literal listener grounding

literal L0-listener

pragmatic L1-speaker

hyper-pragmatic L2-listener

“Inverse-RSA”
literal speaker grounding

literal L0-speaker

pragmatic L1-listener

hyper-pragmatic L2-speaker

Rabin (1990), Franke & Jäger (2014)

“standard RSA”

literal listener grounding

literal L0-listener

$$P_{L_0}(s \mid u) \propto P(s)\ \mathfrak{L}(s,u)$$

pragmatic L1-speaker

$$P_{S_1}(u \mid s) = \mathrm{SM}\left(\log P_{L_0}(s \mid u) - \mathrm{C}(u)\right)$$

hyper-pragmatic L2-listener

$$P_{L_2}(s \mid u) \propto P(s)\ P_{S_1}(u \mid s)$$

“inverse RSA”

literal speaker grounding

literal L0-speaker

$$P_{S_0}(u \mid s) \propto P(u)\ \mathfrak{L}(u,s)$$

pragmatic L1-listener

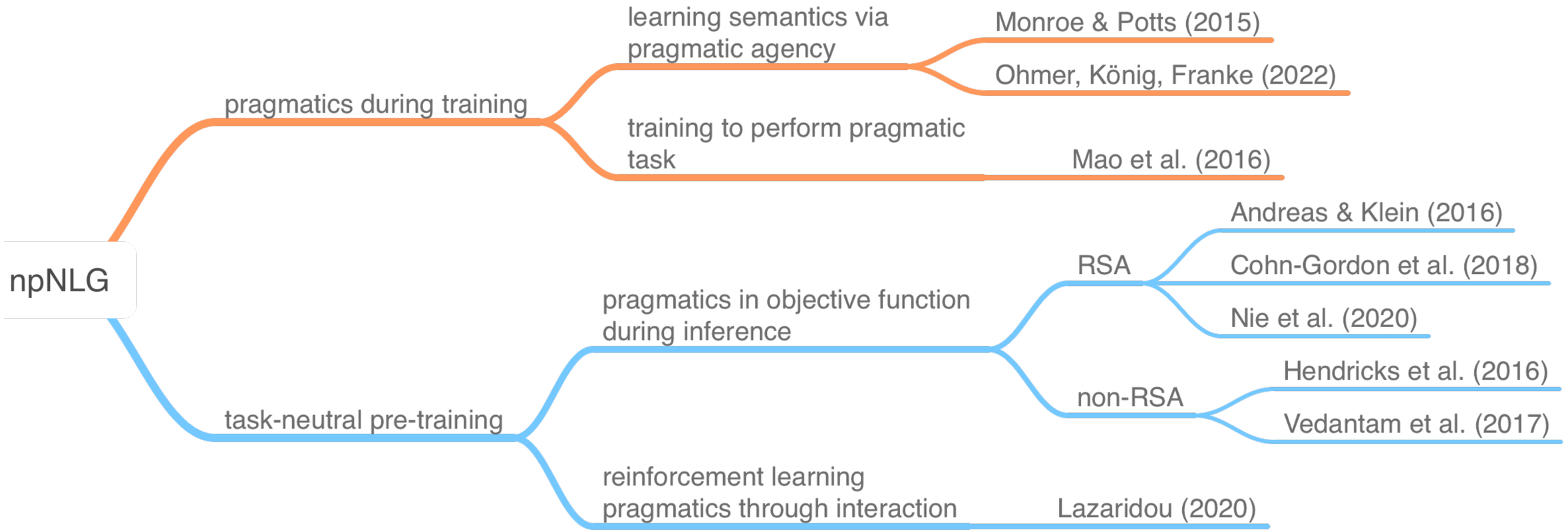
$$P_{L_1}(s \mid u) \propto P(s)\ P_{S_0}(u \mid s)$$

hyper-pragmatic L2-speaker

$$P_{S_2}(u \mid s) = \mathrm{SM}\left(\log P_{L_1}(s \mid u) - \mathrm{C}(u)\right)$$

Overview

different kinds of npNLG approaches



Learning in the RSA model

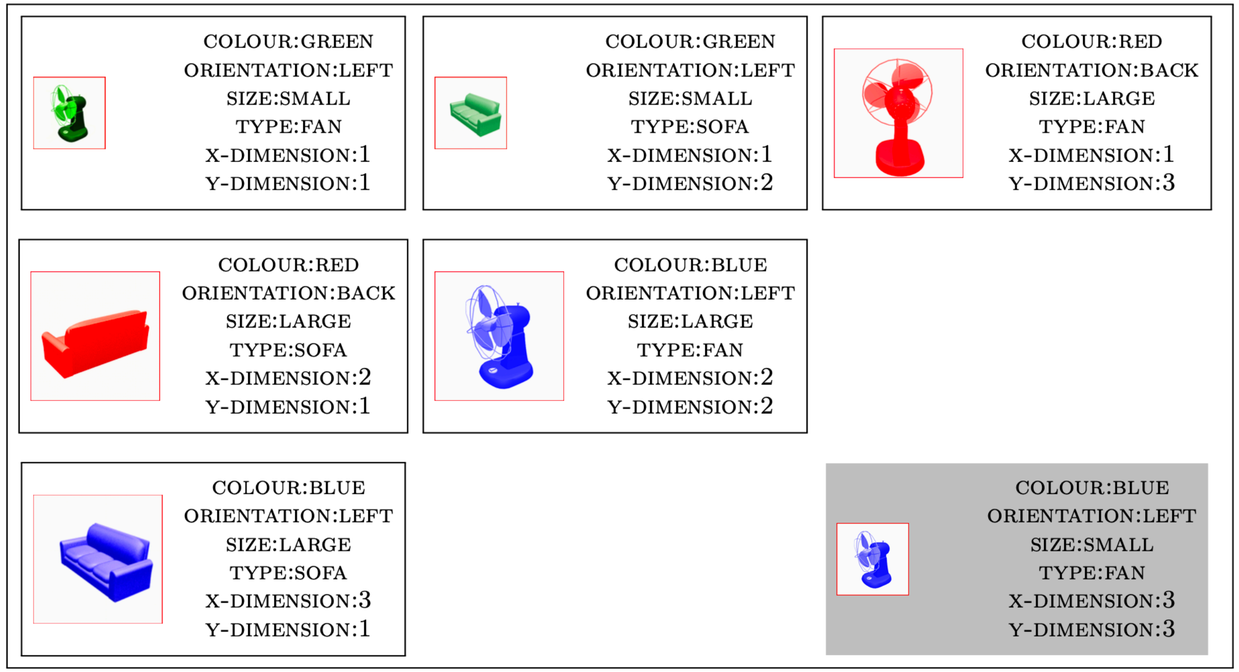
Monroe & Potts (2015), Proc. of Amsterdam Colloquium

Learning in the RSA model

data & modeling set-up

- **goal:** use empirical data to infer semantic meaning that optimizes performance of a speaker model (literal or pragmatic)
- data from TUNA corpus
  - human referential descriptions
  - annotated discrete features of objects
- literal meanings are learned from corpus data
  - $\mathfrak{L}(s, u, c) = \theta^T \varphi(s, u, c)$, where
    - θ^T is a linear mapping
    - $\varphi(s, u, c)$ is a feature representation function
- inverse RSA architecture
  - $P_{S_0}(u \mid s, c) = SM(\mathfrak{L}(s, u, c))$
  - $P_{L_1}(s \mid u, c) \propto P_{S_0}(u \mid s, c)$
  - $P_{S_2}(u \mid s, c) = SM(P_{L_1}(s \mid u, c))$

example from the TUNA corpus



Utterance: "blue fan small"
Utterance attributes: [*colour:blue*]; [*size:small*]; [*type:fan*]

Monroe & Potts (2015)

Learning in the RSA model

evaluation & results

- evaluation metrics:
  - compare features selected by human & machine
  - **accuracy:** perfect match in all features
  - **dice score:** degree of overlap selected features
- models compared:
  - untrained RSA (just using features)
  - speaker models with learned semantics:
    - literal vs pragmatic speakers
    - based on different kinds of features:
      - basic features
      - additional information on human-like generation
- upshot & evaluation:
  - outperforms RSA (w/ predefined meanings)
  - trained S1 is best on aggregate data
  - **BUT:** requires a curated set of discrete features

results reported in the paper

Model	Furniture		People		All	
	Acc.	Dice	Acc.	Dice	Acc.	Dice
RSA s_0 (random true message)	1.0%	.475	0.6%	.125	1.7%	.314
RSA s_1	1.9%	.522	2.5%	.254	2.2%	.386
Learned S_0, basic feats.	16.0%	.779	9.4%	.697	12.9%	.741
Learned S_0, gen. feats. only	5.0%	.788	7.8%	.681	6.3%	.738
Learned S_0, basic + gen. feats.	**28.1%**	**.812**	17.8%	.730	***23.3%***	***.774***
Learned S_1, basic feats.	23.1%	.789	11.9%	.740	17.9%	.766
Learned S_1, gen. feats. only	17.4%	.740	1.9%	.712	10.3%	.727
Learned S_1, basic + gen. feats.	***27.6%***	.788	**22.5%**	**.764**	**25.3%**	**.777**

Monroe & Potts (2015)

Pragmatic Reinforcement Learning

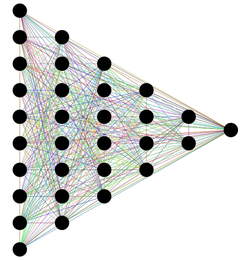
Ohmer, Franke & König (2021), Cognitive Science

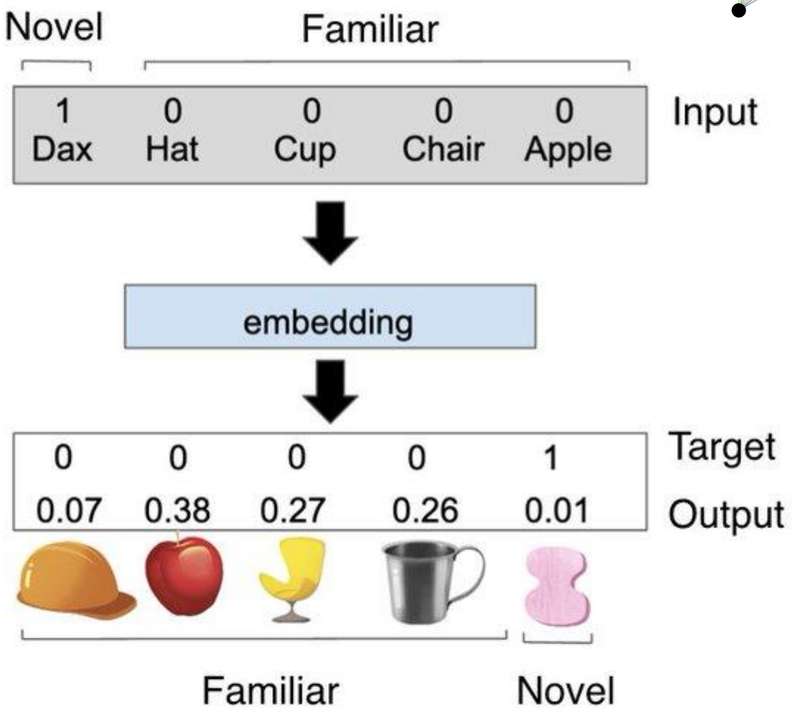
Mutual exclusivity (ME) bias

CSP-Subheading

Anti-ME bias in neural networks

Ghandi & Lake (2020, *arXiv*)

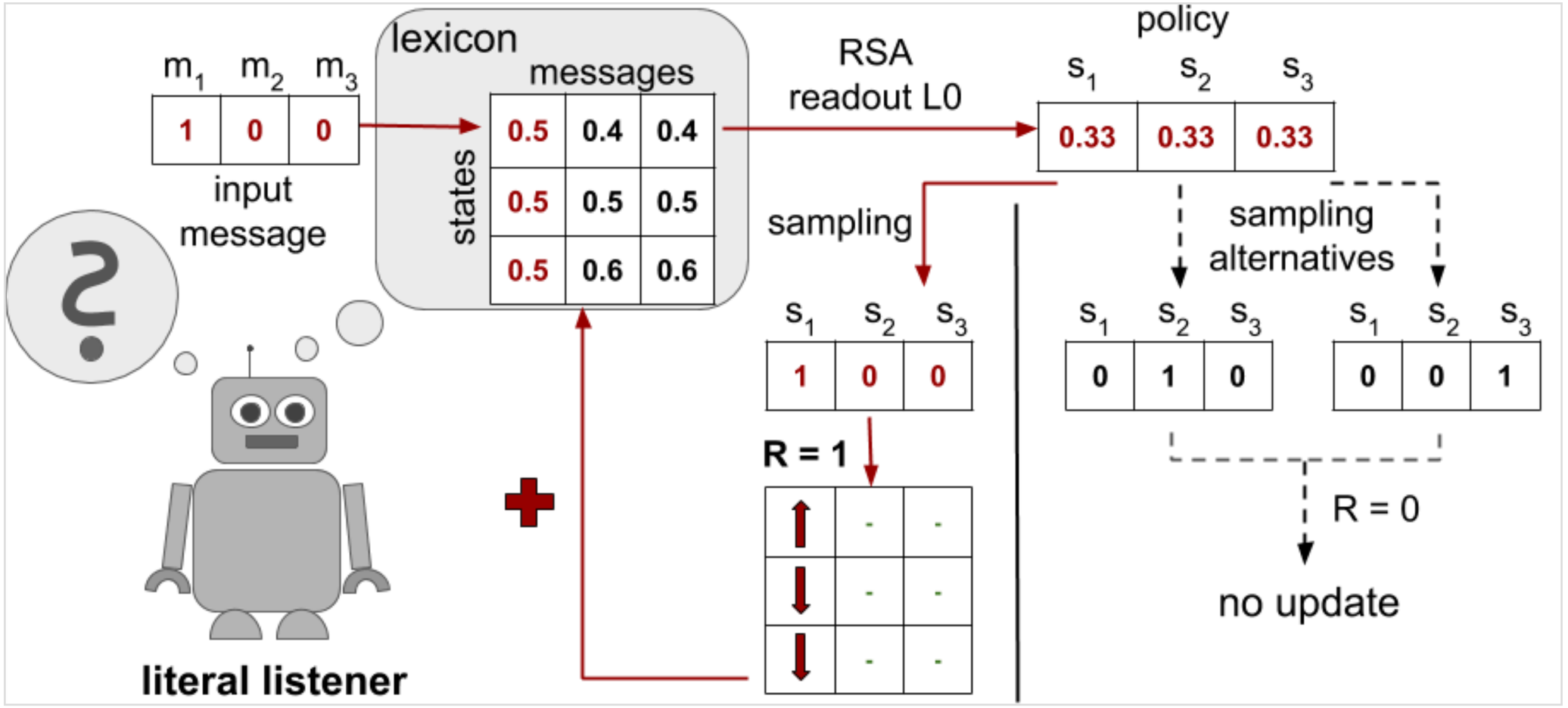




Gradient-based RL of semantic values

literal agents

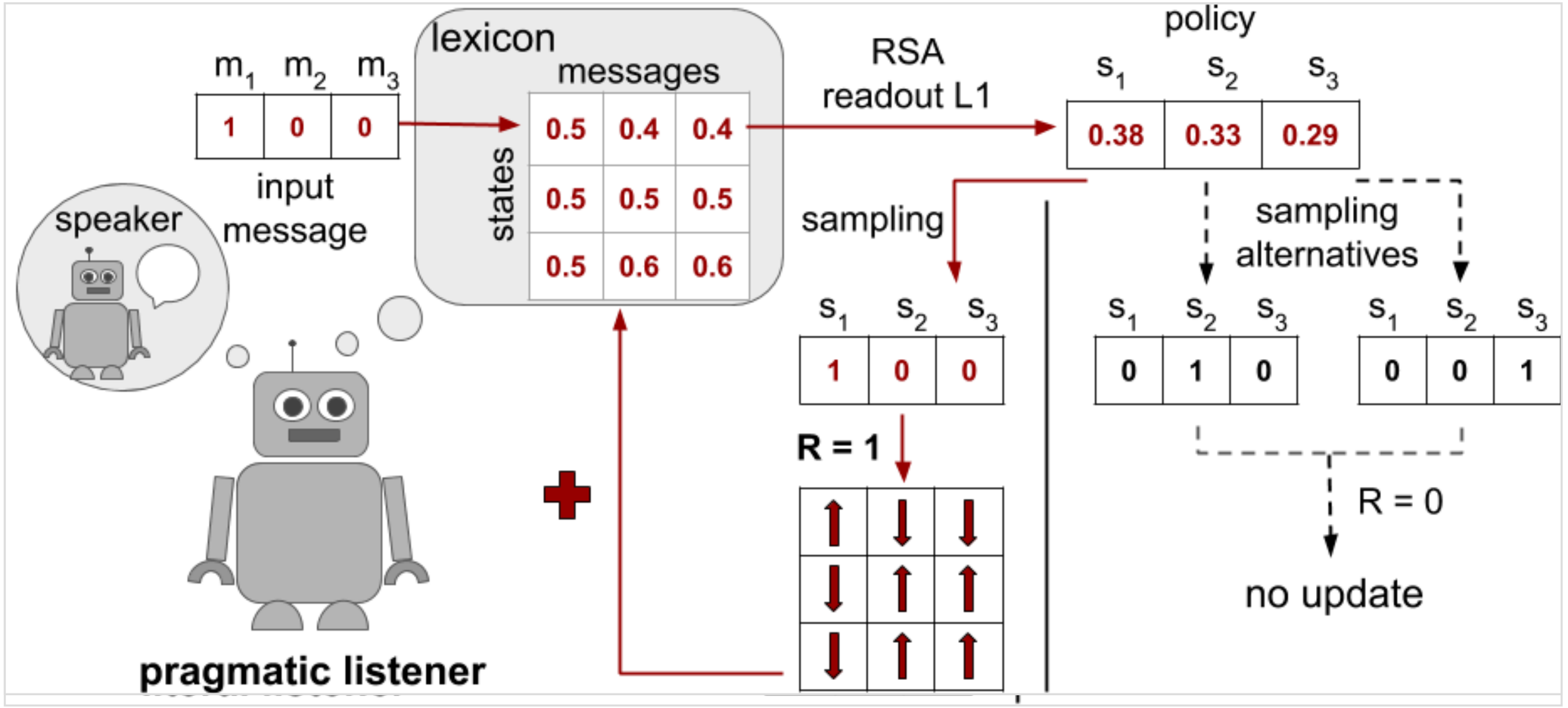
- agents update lexical meanings via RL
- policy defined by lexicon



Gradient-based RL of semantic values

pragmatic agents

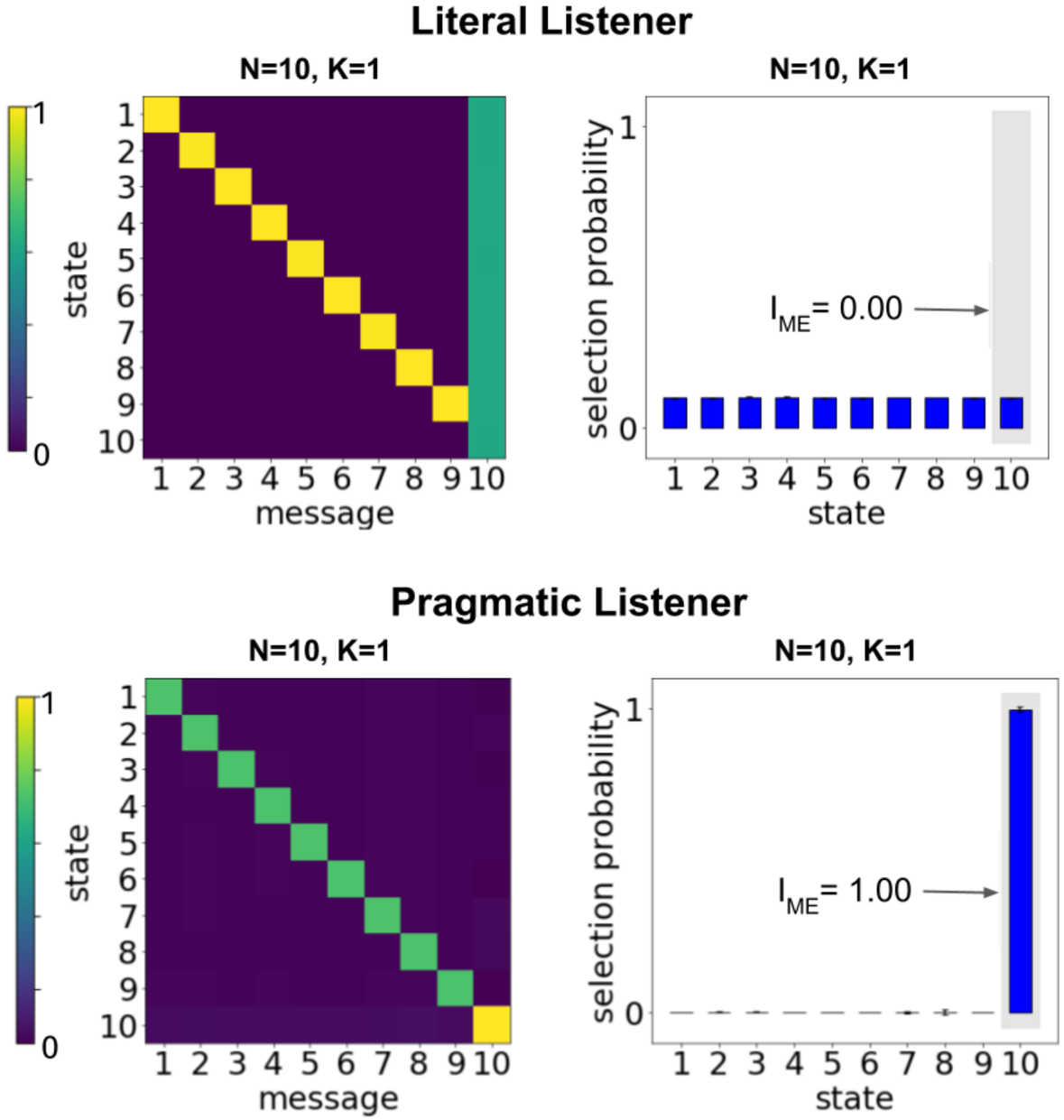
- agents update lexical meanings via RL
- policy defined by lexicon **& RSA**



Simulation set-up & results

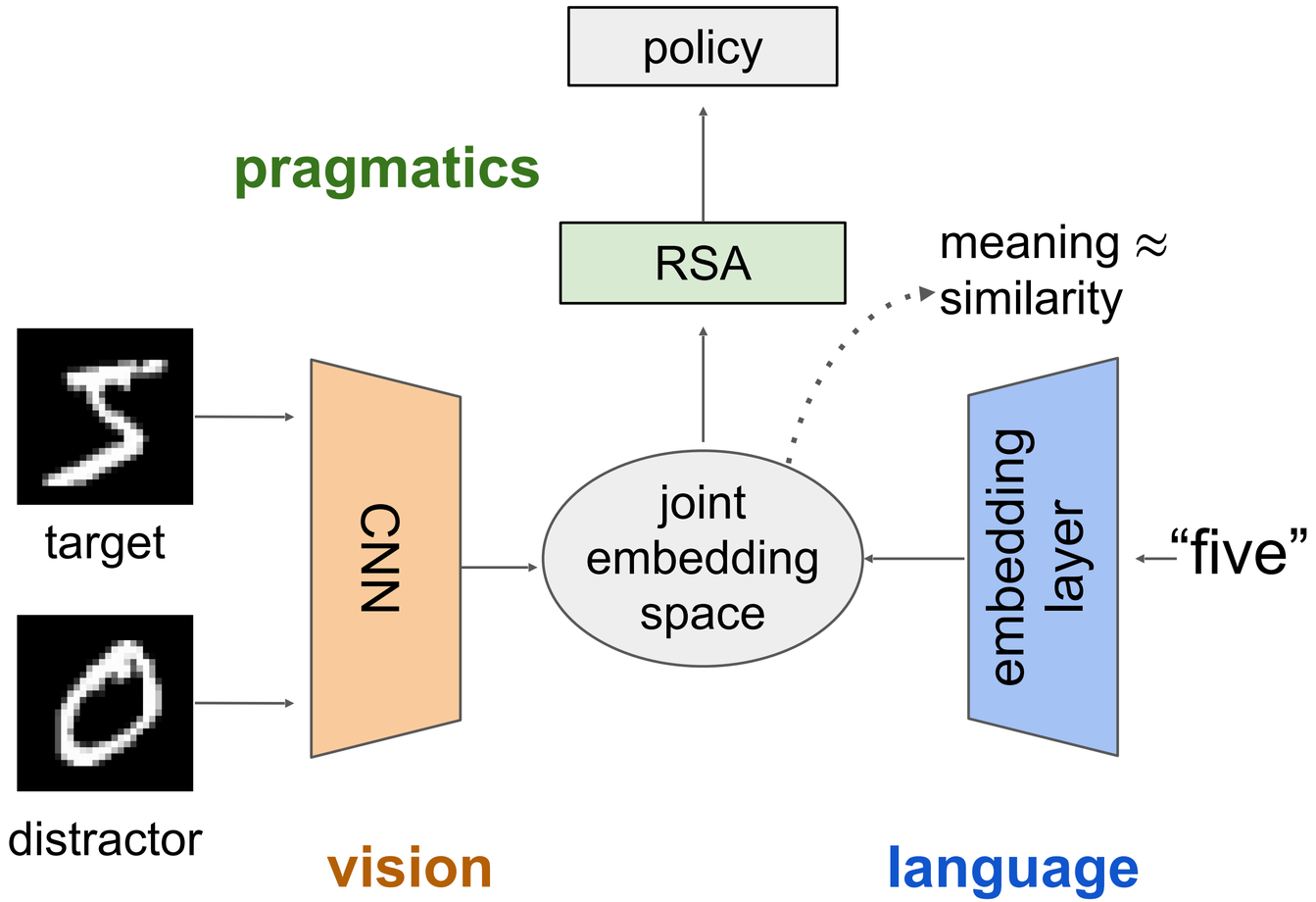
CSP-Subheading

- set-up:
  - 10 states and messages matched 1-to-1
  - 9 pairs for training
  - 1 hold-out pair (index 10) for testing
- results:
  - lexical and behavioral ME bias for pragmatic agents, but not for literal agents
- extensions:
  - dynamically growing lexica
  - similarities to human word learning:
    - ME increases with vocabulary size
    - ME increases with exposure



Pragmatic RL in open-ended message & state spaces

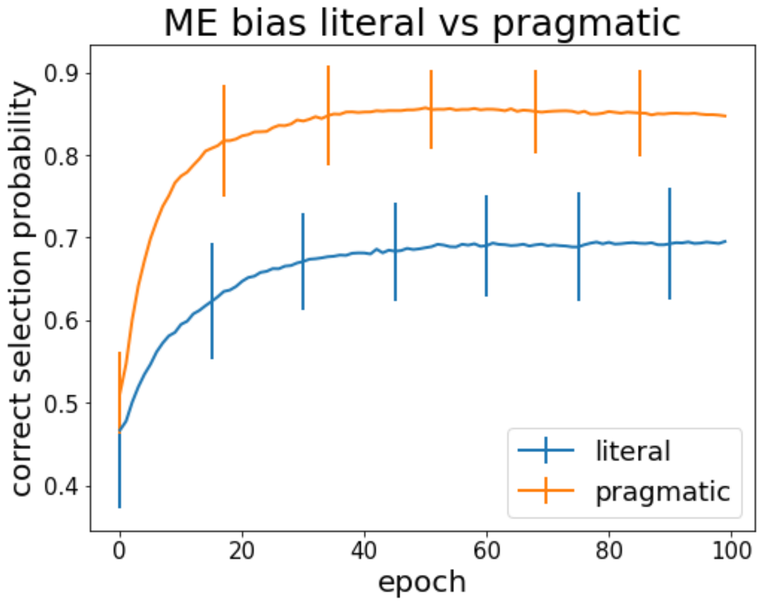
- image embedding
  $f\colon I \to [0;1]^n$
- message embedding
  $g\colon M \to [0;1]^n$
- semantic meaning:
  $\mathfrak{L}(s,m) = f(s) \cdot g(m)$

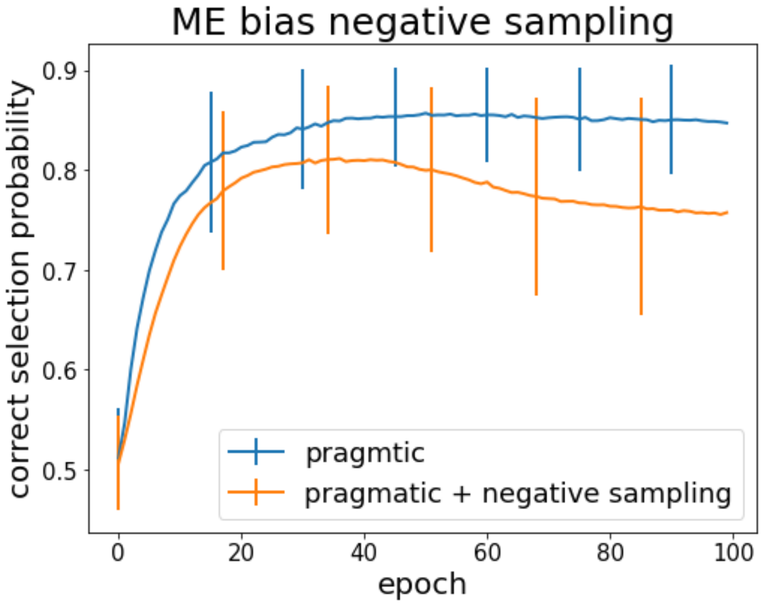


Simulation set-up & results

pragmatic RL w/ joint image-word embeddings

- set-up:
  - MNIST images as states
  - single embedding layer for single-word messages
  - one hold-out state/message
- results:
  - agents show behavioral ME bias
- negative sampling:
  - include non-matching image-word pairs during training marked as “negative examples”
    - Gulordava et al (2020); Vong & Lake (2022)
  - not required w/ pragmatic RL, even detrimental





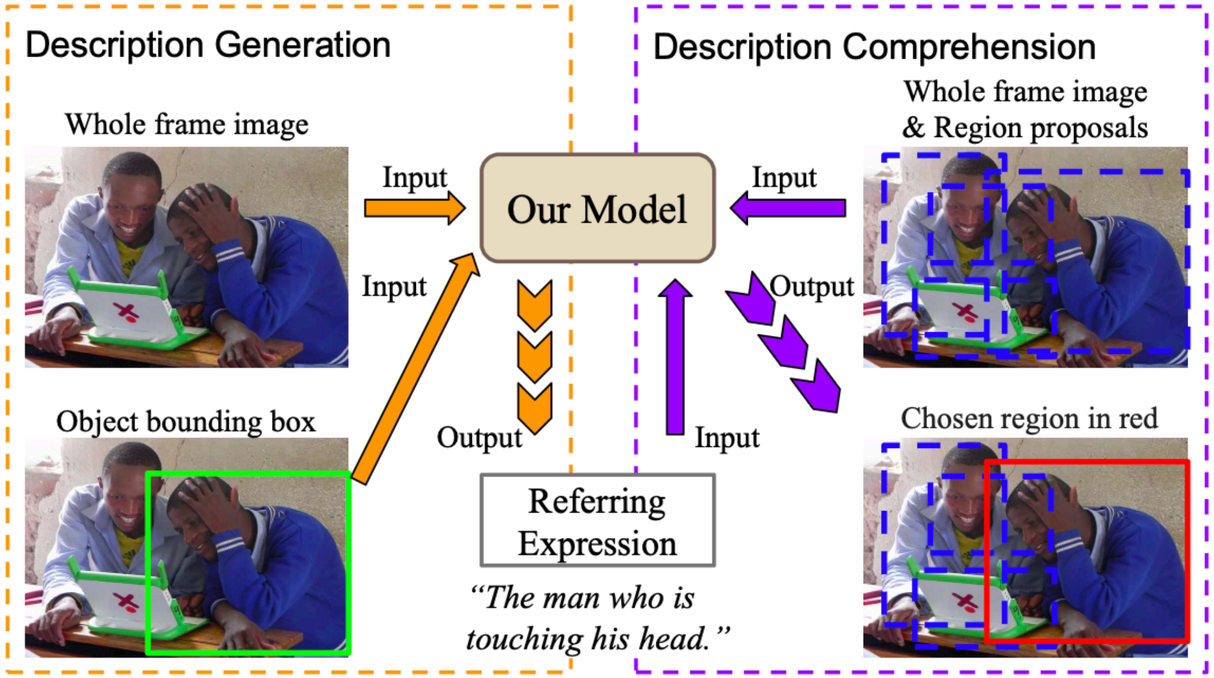
Generation and comprehension of unambiguous object descriptions

Mao et al. (2016), CVPR

Pragmatic object reference

training context-discriminative object descriptions

- **task:**
  - generate (unambiguous) referential description for a target object in an image
  - infer the intended referent object from a given description in an image
- **training set:**
  - Google Refexp data set
  - data points are triples: $\langle c, i, r \rangle$
    - caption
    - image
    - region (bounding box, represents objects)
- **approach:**
  - train S_0 and S_2 from "inverse RSA"



Mao et al. (2016)

Pragmatic object reference

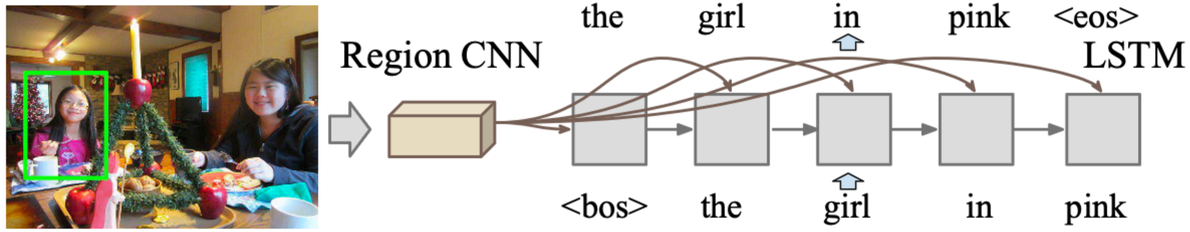
system architecture

- literal speaker:
  - $P_{S_0}(c \mid i, r)$
  - trained as image captioner w/ objective function:

    $-\log P_{S_0}(c \mid i, r)$

- pragmatic listener:
  - $P_{L_1}(r \mid c, i) \propto P_{S_0}(c \mid i, r)$ [uniform priors]
  - implicit competitor set $R(i)$:
    - all objects in the picture
    - all objects of the same category
    - randomly generated bounding boxes
- pragmatic speaker:
  - $P_{S_2}(c \mid i, r) \propto P_{L_1}(r \mid c, i)$ [$\alpha = 1$]
  - trained as image captioner w/ objective function:

    $-\log P_{L_1}(r \mid c, i)$ [max. mutual information]



Mao et al. (2016)

Pragmatic object reference

results

- human raters: percentage of generated descriptions that are at least as good as the description in the data set:
  - 15.9% for S_0
  - 20.4% for S_1

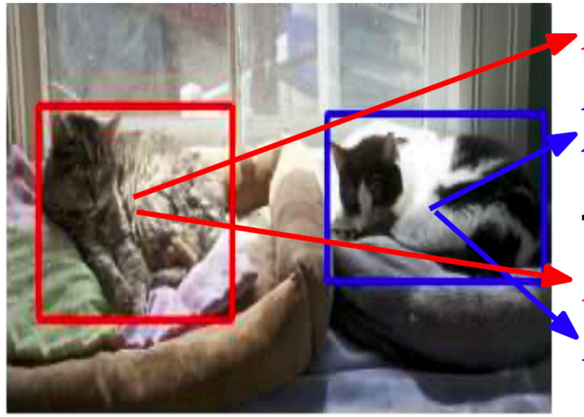
- accuracy of generated descriptions

different competitor sets at test time

	Proposals	GT		Multibox	
	Descriptions	GEN	GT	GEN	GT
S_0	ML (baseline)	0.803	0.654	0.564	0.478
S_2	MMI-MM-easy-GT-neg	0.851	0.677	0.590	0.492
S_2	MMI-MM-hard-GT-neg	**0.857**	**0.699**	0.591	0.503
S_2	MMI-MM-multibox-neg	0.848	0.695	**0.604**	**0.511**
	MMI-SoftMax	0.848	0.689	0.591	0.502

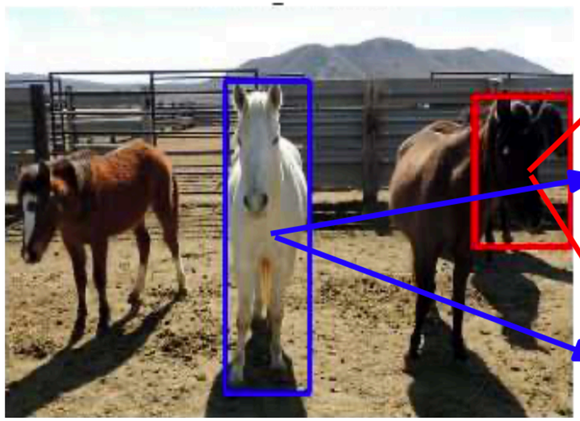
synthetic data

human data



S_2

S_0



S_2

S_0

Mao et al. (2016)

Reasoning about pragmatics w/ neural listeners and speakers

Andreas & Klein (2016), EMNLP

Neural-Pragmatic Natural Language Generation

for contrastive image captioning

- **goal:** produce caption c that picks out target image i_t over distractor i_d
- **data:** image-caption pairs (i_t, c)
- **literal listener:** pre-trained to maximize

  $P_{L_0}(i_t \mid i_t, i_d, c)$ for all pairs (i_t, c)
- **literal speaker:** pre-trained to maximize

  $P_{S_0}(c \mid i_t)$ for all pairs (i_t, c)
- **pragmatic speaker (reranker):**
  - sample candidates:

    $c_1, \ldots, c_n \sim P_{S_0}(\cdot \mid i_t)$
  - score candidates:

    $s_k = P_{L_0}(i_t \mid i_t, i_d, c_k)^{1-\lambda} \; P_{S_0}(c \mid i_t)^{\lambda}$
  - select caption w/ max. score

(a) target

(b) distractor

the owl is sitting in the tree

Andreas & Klein (2016)

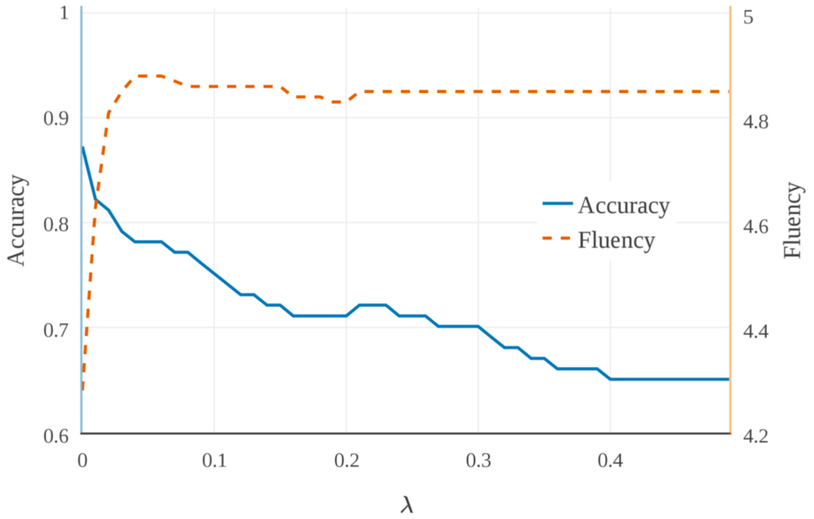
Neural-Pragmatic Natural Language Generation

results

- the more samples we take to score, the higher the accuracy

# samples	1	10	100	1000
Accuracy (%)	66	75	83	85

- accuracy deteriorates with increasing λ



- pragmatic speaker models beats literal speaker baseline, and a reimplementation of the Mao et al. (2015) model

	Dev acc. (%)		Test acc. (%)	
Model	All	Hard	All	Hard
Literal (S0)	66	54	64	53
Contrastive	71	54	69	58
Reasoning (S1)	**83**	**73**	**81**	**68**

Andreas & Klein (2016)

Pragmatically Informative Image Captioning with Character-Level Inference

Cohn-Gordon, Goodman & Potts (2018), NAACL

Incremental neural RSA

model architecture

- **goal:** produce caption c that singles out the target image i_t given a distractor set
- **data:** image-caption pairs (i_t, c)
- **literal speaker:** pre-trained NIC

  $$P_{S_0}(w_{1:n} \mid i)$$ [neural network]
- **L1-listener:** Bayes rule w/ partial captions

  $$P_{L_1}(i \mid w_{1:n}) \propto P_{S_0}(w_{1:n} \mid i)$$ [uniform priors]
- **pragmatic speaker (incremental RSA):**

  $$P_{S_2}(w_{n+1} \mid i, w_{1:n}) \propto P_{S_0}(w_{1:n} \mid i) \;\; P_{L_1}(i \mid w_{1:n})^{\alpha}$$
- **granularity:**
  - word-level: each w_n is a full word
  - character-level: each w_n is a single character

S_0 caption: a double decker bus
S_2 caption: a red double decker bus

Cohn-Gordon, Goodman & Potts (2018)

Excursion

formal details of incremental RSA

$$P_{L_0}(i \mid w_{1:n}) = \frac{P(i) \;\; P_{S_0}(w_{1:n} \mid i)}{\sum_j P(j) \;\; P_{S_0}(w_{1:n} \mid j)}$$ [our reformulation]

$$= \frac{P(i) \;\; P_{S_0}(w_{1:(n-1)} \mid i) \; P_{S_0}(w_n \mid w_{1:(n-1)}, i)}{\sum_j P(j) \;\; P_{S_0}(w_{1:(n-1)} \mid j) \; P_{S_0}(w_n \mid w_{1:(n-1)}, j)}$$ [chain rule]

$$= \frac{\frac{1}{C} P(i) \;\; P_{S_0}(w_{1:(n-1)} \mid i) \; P_{S_0}(w_n \mid w_{1:(n-1)}, i)}{\sum_j \frac{1}{C} P(j) \;\; P_{S_0}(w_{1:(n-1)} \mid j) \; P_{S_0}(w_n \mid w_{1:(n-1)}, j)}$$ [introducing constant]

$$= \frac{\frac{P(i) \;\; P_{S_0}(w_{1:(n-1)} \mid i)}{\sum_k P(k) \;\; P_{S_0}(w_{1:(n-1)} \mid k)} \; P_{S_0}(w_n \mid w_{1:(n-1)}, i)}{\sum_j \frac{P(j) \;\; P_{S_0}(w_{1:(n-1)} \mid j)}{\sum_k P(k) \;\; P_{S_0}(w_{1:(n-1)} \mid k)} \; P_{S_0}(w_n \mid w_{1:(n-1)}, j)}$$ [set k to normalization term]

$$= \frac{P(i \mid w_{1:n}) \; P_{S_0}(w_n \mid w_{1:(n-1)}, i)}{\sum_j P(j \mid w_{1:n}) \; P_{S_0}(w_n \mid w_{1:(n-1)}, j)}$$ [formulation from the paper]

Excursion

formal details of incremental RSA

$$P_{S_1}(w_{n+1} \mid i, w_{1:n}) \propto \exp\left(\alpha \left(\log P_{L_0}(i \mid w_{1:n}) - \text{Cost(u , i)}\right)\right)$$

[vanilla RSA

$$\propto P_{L_0}(i \mid w_{1:n})^{\alpha} \ \exp\left(-\text{Cost(u , i)}\right)$$

[rules of exponential function]

$$= P_{S_0}(w_{1:n} \mid i) \ \ P_{L_0}(i \mid w_{1:n})^{\alpha}$$

[defining costs as S_0 production]

Upshot:

incremental RSA is, by definition, just plan vanilla RSA
(with a special interpretation of the cost term)

Incremental neural RSA

results

- compare literal and pragmatic models, for character- and word-level incremental predictions
  - but table shows possibly misleading contrast
  - Char S_2 uses beam search for decoding (beam size 10) but Word S_2 uses greedy decoding
  - with greedy decoding Char S_2 scores 61.2% on TS1
  - the advantage could solely come from different decoding

Model	TS1	TS2
Char S_0	48.9	47.5
Char S_1	**68.0**	**65.9**
Word S_0	57.6	53.4
Word S_1	60.6	57.6

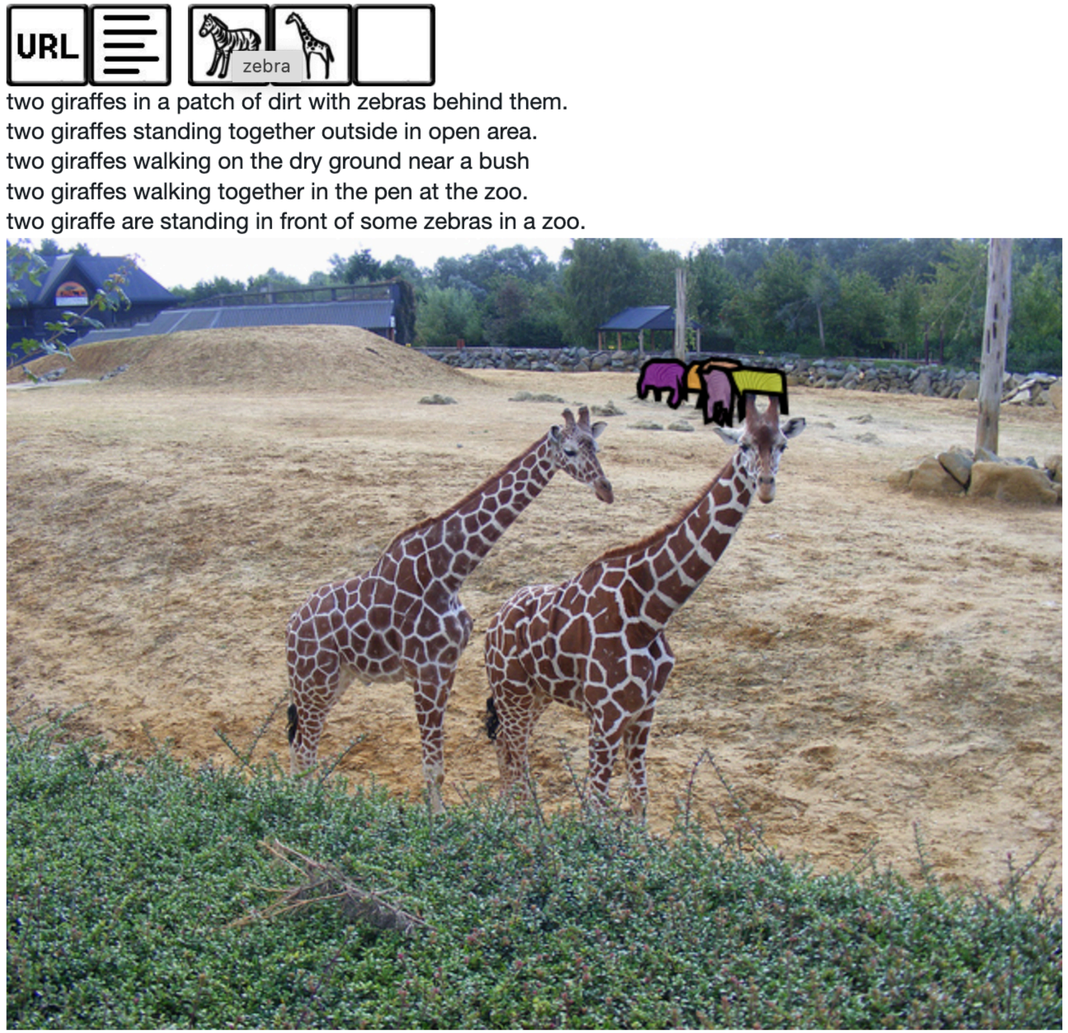
Cohn-Gordon, Goodman & Potts (2018)

Data Sets

MSCOCO

large data set w/ images, captions & labelled-objects

- > 300k images with:
  - captions
  - bounding boxes for 80 objects w/ labels
    - things (concrete objects) and stuff (background elements)
- **URL:** https://cocodataset.org



Yin et al. (2014), “Microsoft COCO: Common Objects in Context”, ECCV

Google Refexp

referential expressions for objects in MS-COCO images

- subset of images from MS-COCO w/ additional referential expressions for objects in the images
- > 26k images with 54k target objects
  - each object types occurs 2-4 times in the picture
  - all objects of that type are sufficiently salient
  - bounding boxes and labels for objects (from MS-COCO)
- ~1.9 referential expressions per target object
  - obtained from MTurk human annotation
    - human producer types referential expression E
    - human interpreter tries to identify target object based on E
    - if successful E is added to data set, if not discarded
- **URL:** Google Refexp

The black and yellow backpack sitting on top of a suitcase.

A yellow and black back pack sitting on top of a blue suitcase.

An apple desktop computer.

The white IMac computer that is also turned on.

Mao et al. (2016) "Generation and Comprehension of Unambiguous Object Descriptions" , CVPR