Concepts, Composition, and Conversational Coordination

Semantic Competence for Situated Interaction

4th Seminar: Concepts (Part III), Composition, Coordination

David Schlangen
University of Potsdam, Germany

http://clp.ling.uni-potsdam.de

https://github.com/davidschlangen/cosine-paris

functional *reprsnt.nal* *composition* *coordination*

Conceptual Apparatus

Inference

- word to word
- discourse resolution

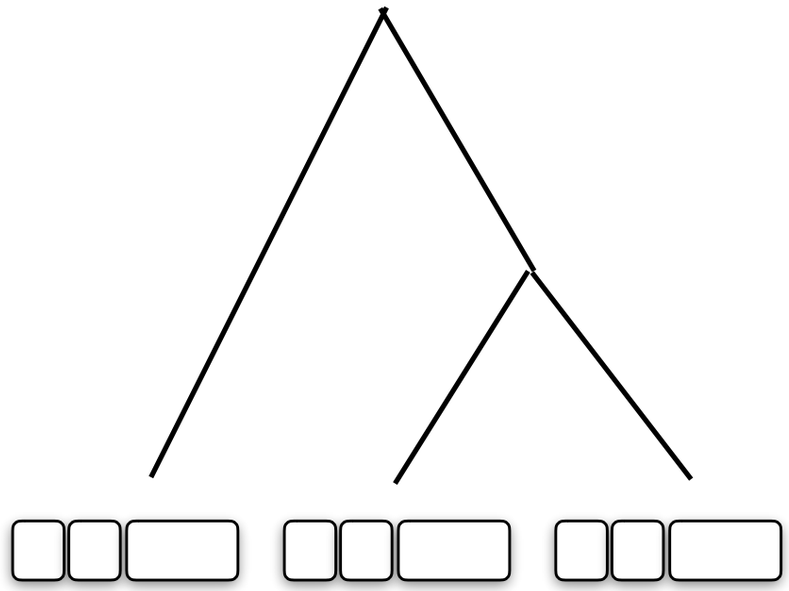
Reference

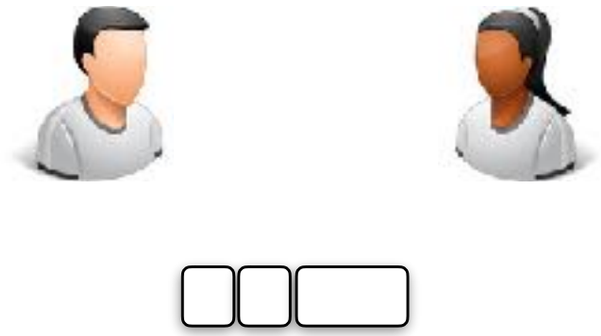
- word to world
- categori-sation
- naming / resolution

symbolic repr.

continuous repr.

classifiers on perceptual input





Look at the white dog!

We just saw a cute dog.

The cutest poodle ever!

Actually, that wasn't a poodle. It was too tall. It was a labradoodle.

- learning
  - incremental (within concept, within vocab)
  - fast
- implemented & tested on real data

functional *reprsnt.nal*

Conceptual Apparatus

Inference

- word to word
- discourse resolution

Reference

- word to world
- categori-sation
- naming / resolution

symbolic repr.

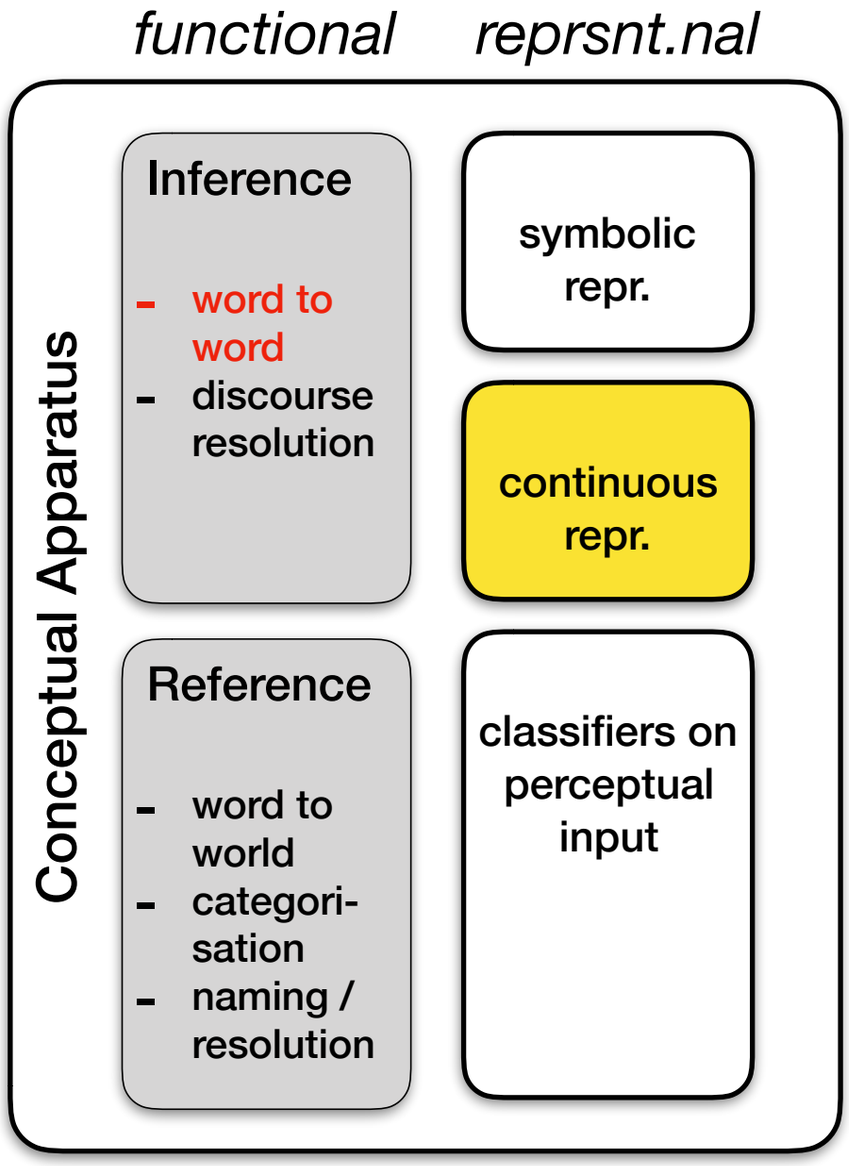
continuous repr.

classifiers on perceptual input

Look at the white dog!

We just saw a cute dog.

The cutest poodle ever!



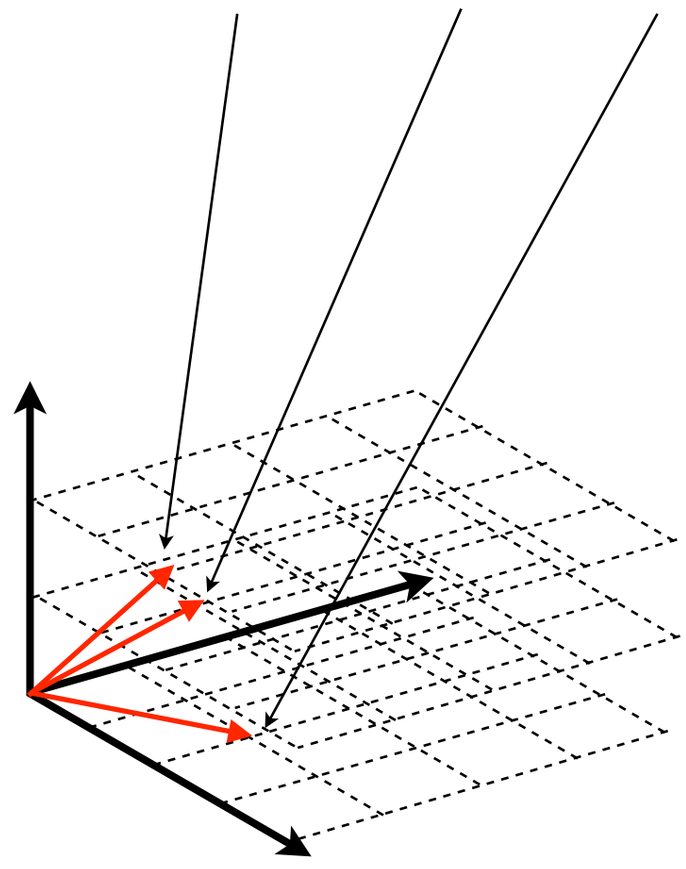
Harris (1954): “If A and B have almost identical environments we say that they are synonyms.”

Firth (1957): “You shall know a word by the company it keeps!”

What is an “environment”, and what is “company” for a word?

Just the surrounding text, or more of the surrounding context?

	Doc1	Doc2	Doc3
Bielefeld	2	1	2
Kern	0	1	1
Universität	2	1	0



structured contexts

continuous repr.

embeddings, from different kinds of context:

- **ref.exp. as sentence, whole corpus, v_{txt}**
- **co-referential exp. as context, v_{ref}**
- **situation as context, v_{sit}**

(Zarrieß & Schlangen, EMNLP 2017)

today

- learning word meaning representations from observations of linguistic contexts
  - different kinds of contexts
  - inference: referential compatiblity
    - predicting antonymy
- cross-over: zero-shot learning
- composition: syntax as interface
  - a side note: use visual denotations to induce structure?
  - composing continuous representations for inference: LSTMs, Transformers, Tree-LSTMs
  - composing references

evaluating the continuous representations

- question: can we get word meaning representations from these kinds of contexts that are useful for our “inference” tasks?

- v_{sit}, v_{ref}, v_{txt} trained with word2vec (gensim implementation), 300 dim, CBOW, on training section of refcoco, refococ+, grex (~ 1 million tokens)

- for comparison, 300 dim w2v pre-trained on 3 billion tokens (GoogleNews), v_{gn}

- NB: word2vec in its original form is non-incremental, doesn’t handle extension of vocab. There are attempts to overcome that (e.g., Kaji & Kobayashi 2017), but no real focus of research.

task 1: predict human similarity ratings

Method:

- correlation of model prediction with human judgement (Spearman's rho)

Datasets:

- "*compatibility*" (Kruszewski & Baroni 2015), "how well could something referred to with the one term also be referred to with the other?"

- "SL-sim-vis/sem" (Silberer & Lapata 2014), selected noun pairs from McRae *et al.* (2005) norms ("Participants were asked to rate a pair on two dimensions, visual and semantic similarity using a Likert scale of 1 (highly dissimilar) to 5 (highly similar)."

- "*MEN*" (Bruni *et al.* 2012), 3,000 pairs of words that occured as ESP game tags

task 1: predict human similarity ratings

	compat	SL-vis	SL-sem	MEN
V_{sit}	0.21	0.37	0.41	0.42
V_{txt}	0.18	0.24	0.29	0.23
V_{ref}	0.23	0.36	0.40	0.44
V_{gn}	**0.26**	**0.57**	**0.73**	**0.77**

task 2: name / name

- premise: a single noun (= head noun taken from referring expression)

- hypothesis: another single noun, either from same region, or from another

- question: can the hypothesis be used in reference to the same object as the premise?

- data: validation sections of corpora (to train model), test section

```
((‘guy', 'man'), 495),
(('girl', 'woman'), 234),
(('lady', 'woman'), 210),
(('shirt', 'man'), 173),
(('person', 'woman'), 172),
(('person', 'man'), 159),
(('person', 'guy'), 130),
(('man', 'person'), 122),
(('man', 'guy'), 117),
(('shirt', 'woman'), 116),
(('shirt', 'guy'), 116),
(('shirt', 'person'), 112),
(('man', 'shirt'), 104),
(('lady', 'girl'), 102),
(('guy', 'shirt'), 99),
(('guy', 'person'), 98),
(('kid', 'boy'), 92),
(('girl', 'person'), 80),
(('lady', 'person'), 68),
(('person', 'shirt'), 61),
(('motorcycle', 'bike'), 59),
(('dude', 'man'), 59),
(('girl', 'shirt'), 57),
(('shirt', 'boy'), 55),
(('player', 'guy'), 47),
(('kid', 'child'), 42)
```

```
('ship', 'cupcake'),
('canoe', 'bird'),
('salad', 'cot'),
('row', 'trailer'),
('buildings', 'choc'),
('blade', 'momma'),
('dude', 't-shirt'),
('plants', 'tvs'),
('statues', 'passenger'),
('hills', 'broccoli'),
('racket', 'messenger'),
('picture', 'area'),
('lots', 'right-hand'),
('bottle', 'bit'),
('vulture', 'total'),
('metal', 'kitchen'),
('door', 'horse'),
('zebra', 'ladie'),
('yup', 'statue'),
('pice', 'bathtub'),
('rigth', 'veh'),
('elephant', 'giraffes'),
('senator', ‘stylist'),
(‘cup', 'stripe'),
('surf', 'sun'),
```

task 2: name / name

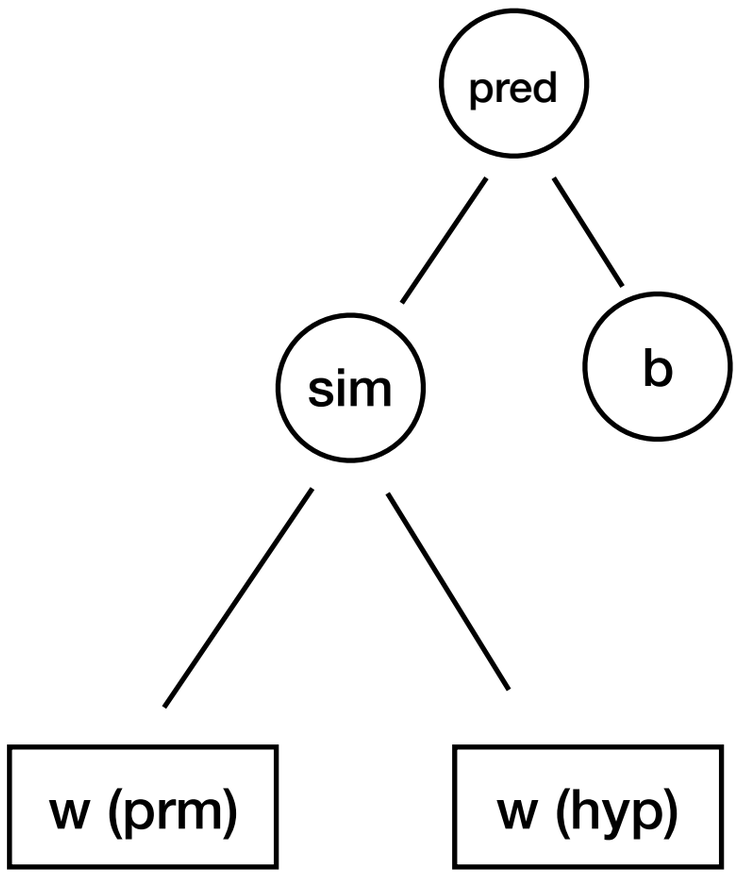
- premise: a single noun (= head noun taken from referring expression)

- hypothesis: another single noun, either from same region, or from another

- question: can the hypothesis be used in reference to the same object as the premise?

- data: validation sections of corpora (to train model), test section

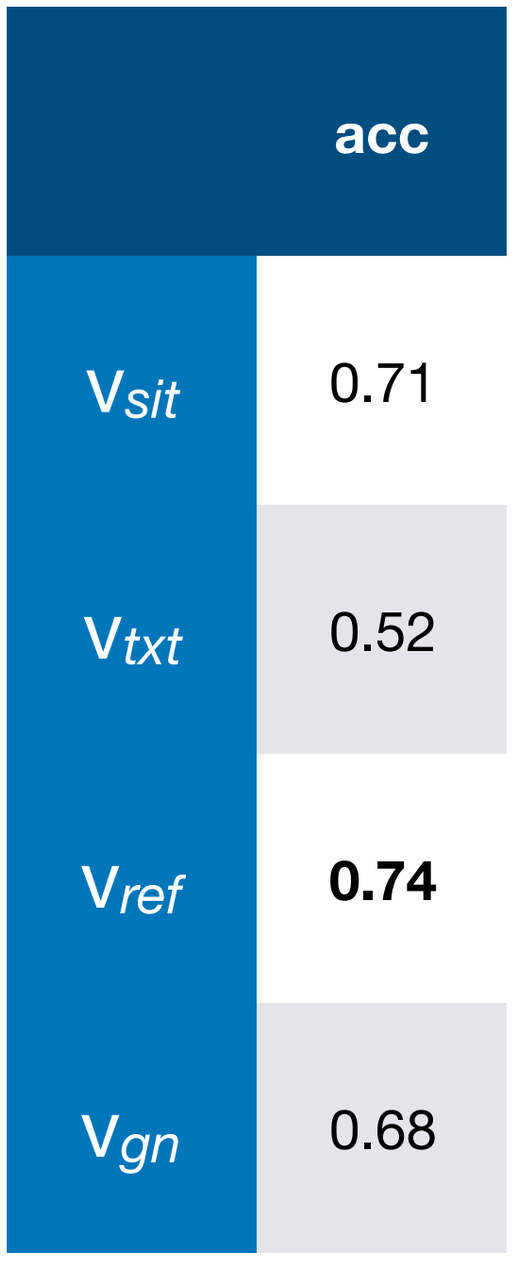
Model:



Logistic regression learns threshold on similarity above which to predict “entails”

task 2: name / name

- premise: a single noun (= head noun taken from referring expression)
- hypothesis: another single noun, either from same region, or from another
- question: can the hypothesis be used in reference to the same object as the premise?
- data: validation sections of corpora (to train model), test section



task 3: rex / rex

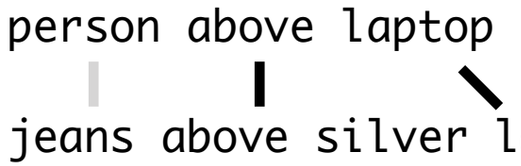
- premise: a referring expression
- hypothesis: another referring expression, from same image
- question: can the hypothesis be used in reference to the same object as the premise?
- data: validation sections of corpora (to train model), test section

```
neutral: person on right reaching out || left guy
neutral: person on right reaching out || man white shirt back
neutral: person right || man white shirt back
entailment: white shirt || man white shirt back
neutral: white shirt || woman in purple
entailment: man on left || old guy
entailment: man on left || left guy
entailment: old guy || left guy
neutral: man on left || man white shirt back
neutral: man on left || woman in purple
neutral: old guy || white shirt
entailment: person above laptop || jeans above silver laptop
entailment: person above laptop || jeans
entailment: jeans above silver laptop || jeans
neutral: person above laptop || laptop with green keyboard
neutral: person above laptop || jeans on right
neutral: checkered tablecloth || olive on pizza in front
neutral: checkered table cloth || right pizza
entailment: serving surface || table that the food is on
neutral: left bear || right bear
```

task 3: rex / rex

- premise: a referring expression
- hypothesis: another referring expression, from same image
- question: can the hypothesis be used in reference to the same object as the premise?
- data: validation sections of corpora (to train model), test section

"Model":
Try to align the words, finding word pairs with high similarity, add & normalise to get score for phrase pair.



Learn decision model as before.

	acc
V_{ref}	0.64
V_{gn}	0.61
bsln	**0.66**

task 3: rex / rex

- premise: a referring expression
- hypothesis: another referring expression, from same image
- question: can the hypothesis be used in reference to the same object as the premise?
- data: validation sections of corpora (to train model), test section

```
neutral: person on right reaching out || left guy
neutral: person on right reaching out || man white shirt back
neutral: person right || man white shirt back
entailment: white shirt || man white shirt back
neutral: white shirt || woman in purple
entailment: man on left || old guy
entailment: man on left || left guy
entailment: old guy || left guy
neutral: man on left || man white shirt back
neutral: man on left || woman in purple
neutral: old guy || white shirt
entailment: person above laptop || jeans above silver laptop
entailment: person above laptop || jeans
entailment: jeans above silver laptop || jeans
neutral: person above laptop || laptop with green keyboard
neutral: person above laptop || jeans on right
neutral: checkered tablecloth || olive on pizza in front
neutral: checkered table cloth || right pizza
entailment: serving surface || table that the food is on
neutral: left bear || right bear
```

task 4: find “antonyms”

- we have different notions of “contextual similarity”
  - purely textual: these words occur in similar contexts = next to similar words
  - referential: these words occur in references to the same object
- can we bring them together?
- find pairs of words that are similar (text-) contextually, but dissimilar in their reference context
- restricted to third target word, in two word contexts:
  $\{ (w_A, w_B) \mid w_A w_T \in O, w_B w_T \in O, sim_{txt}(w_A,w_A) > \alpha, sim_{ref}(w_A,w_A) < \beta \}$

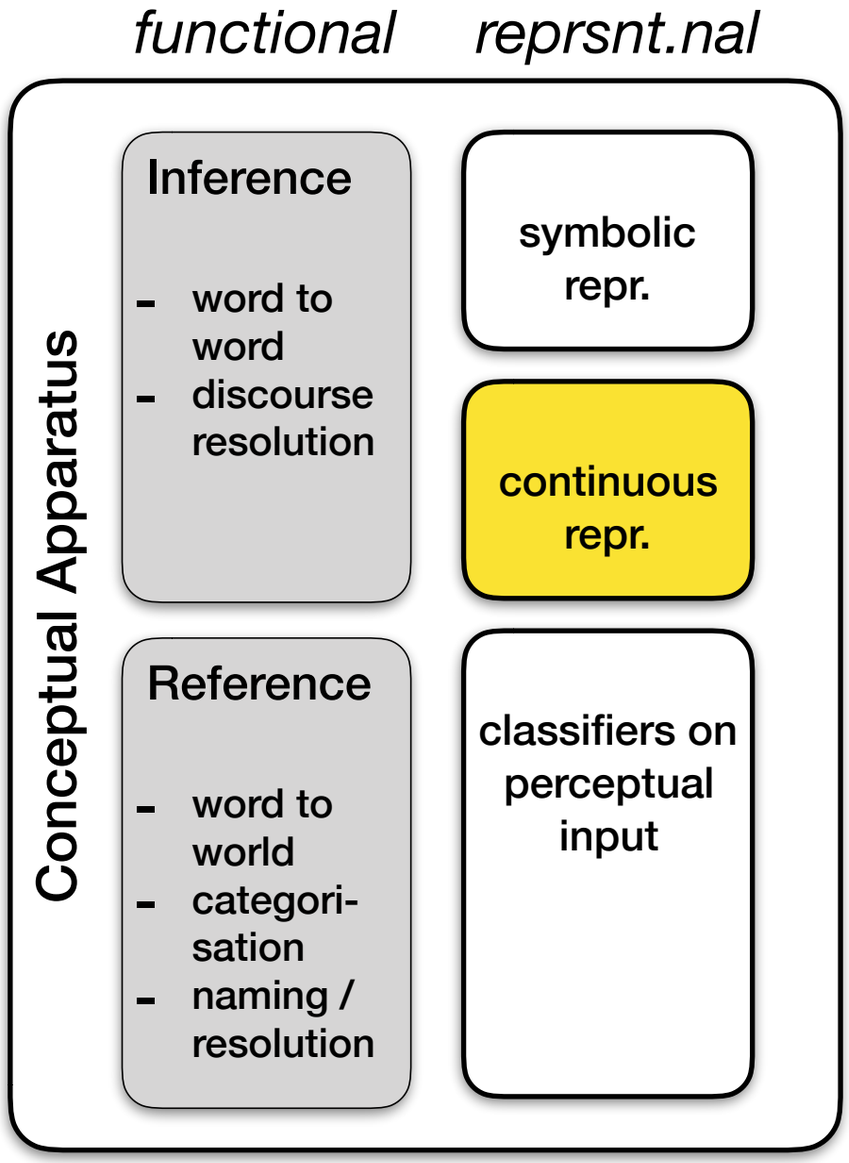
task 4: find “antonyms”

```
***** cat *****
     black     orange
     right       left
***** bench *****
     right       left
***** truck *****
      blue        red
      blue      green
     white        red
     white      green
       red      green
***** couch *****
      left      right
     green      white
     green      black
     green        red
     white        red
     black        red
***** animal *****
      left      right
***** bird *****
       top     bottom
      left      right
```

```
***** giraffe *****
     front       back
      back     middle
     small     middle
     small        big
     small    tallest
     small     taller
    middle        big
    middle    tallest
    middle     taller
      left      right
       big    tallest
       big     taller


***** man *****
      left      right
      tall     center
      tall     middle
 shirtless        old
```

How to evaluate? WordNet has only few of these as antonyms.

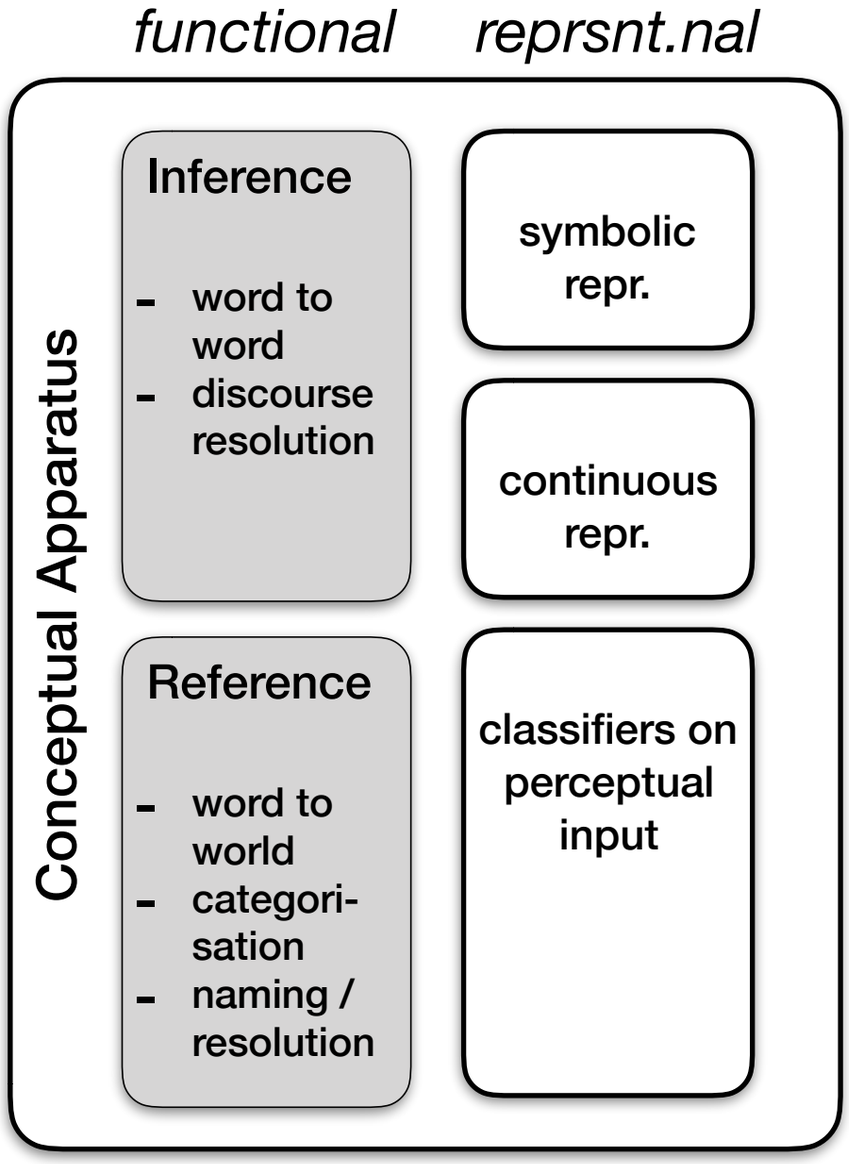


Intermediate Summary

We can derive continuous representations from structured and unstructured (interpreted & uninterpreted) contexts.

These seem to be modelling semantic similarity relations that we would need for our inference tasks.

But the more interesting tasks are not word/word tasks…



Intermediate Summary

Learning situations we have encountered:

- fully interpreted to referent in visual world → perceptual classifiers
- noticing *that* reference was to same object, or within same situation (but not representing percept. info) → v_{ref} , v_{sit}
- just (over)hearing utterances, no attempt at interpretation → v_{txt}

functional *reprsnt.nal*

Conceptual Apparatus

Inference

- word to word
- discourse resolution

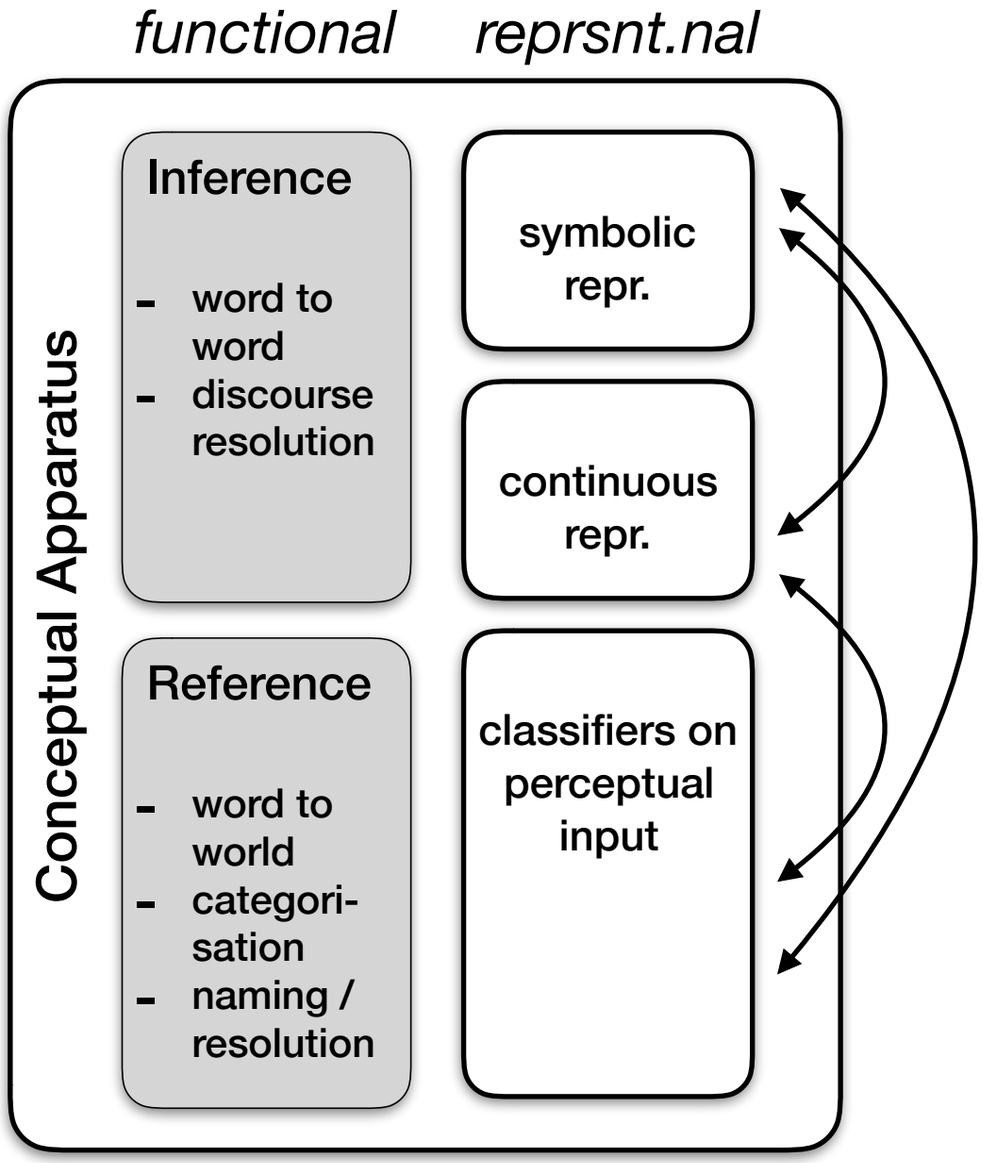
Reference

- word to world
- categori-sation
- naming / resolution

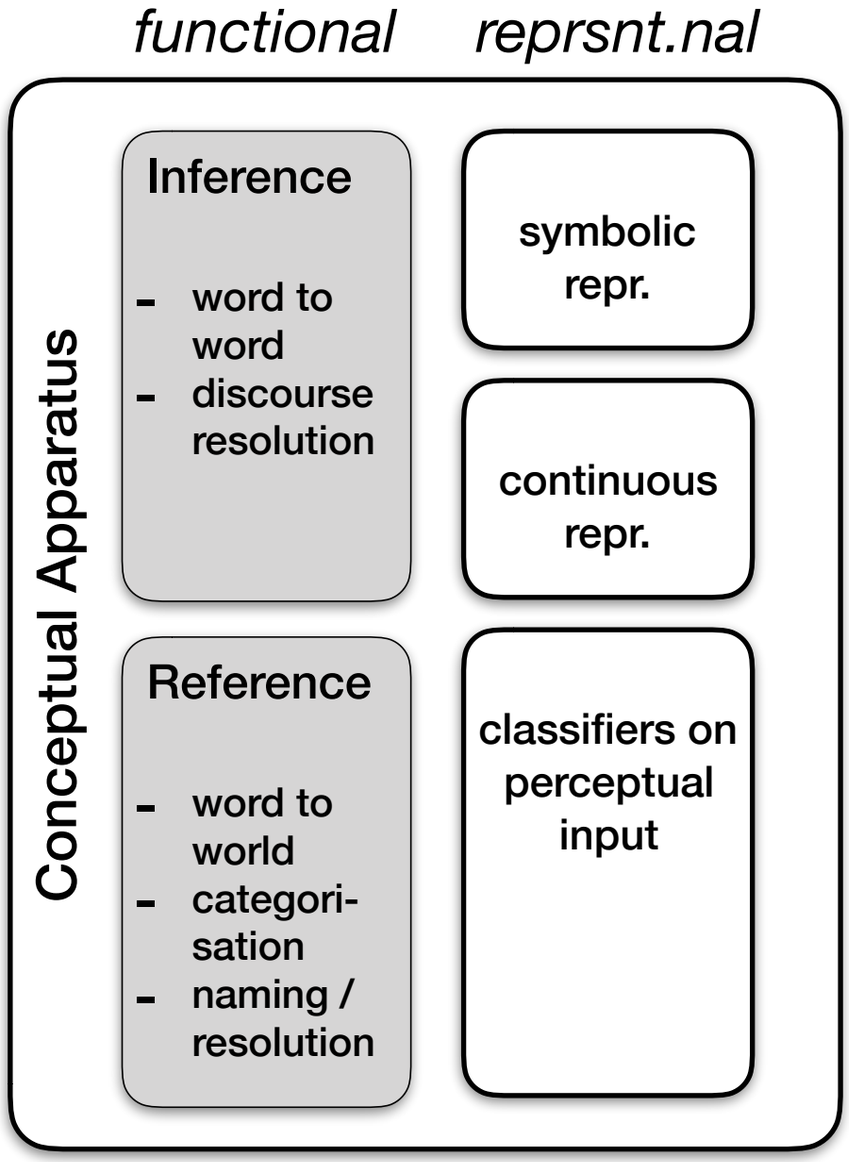
symbolic repr.

continuous repr.

classifiers on perceptual input

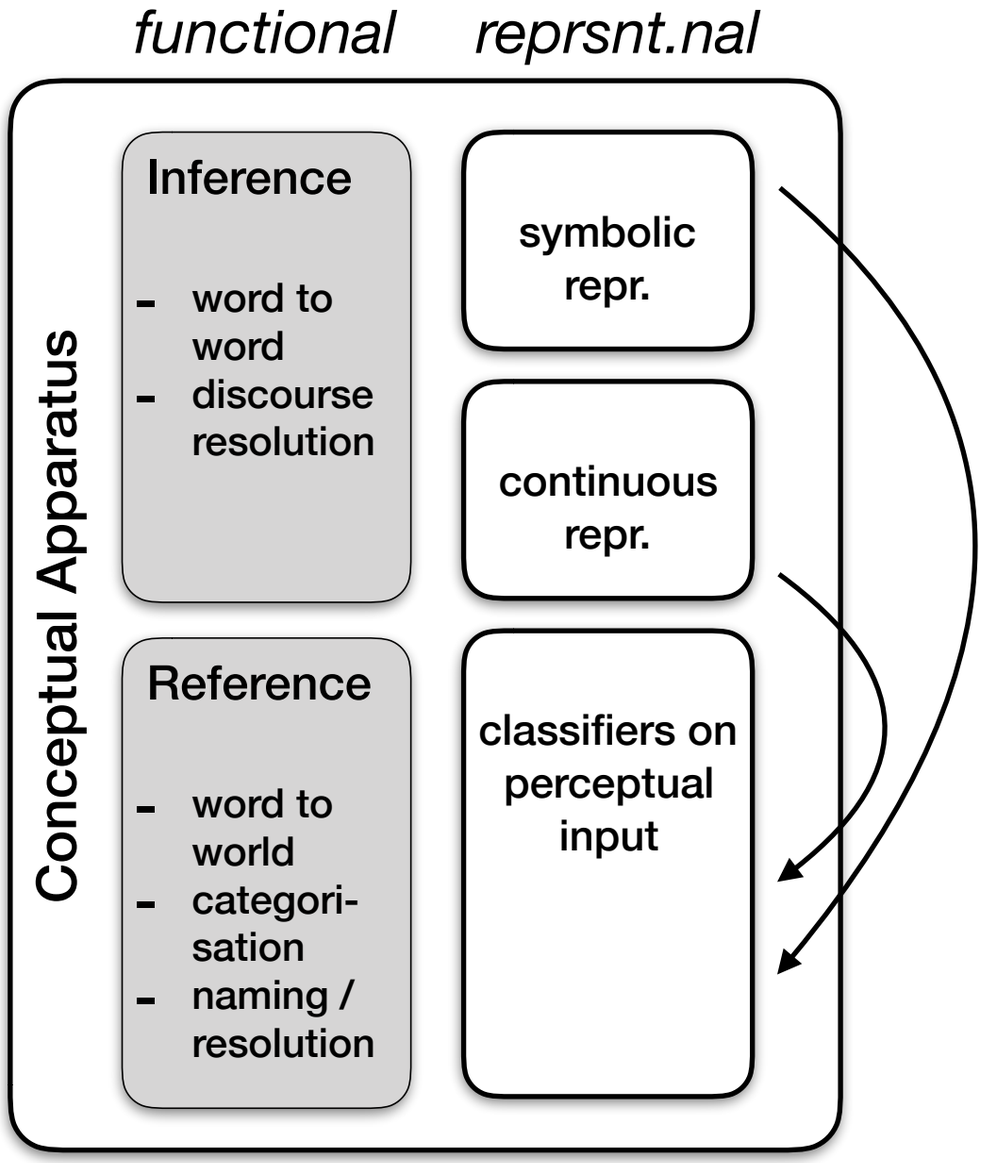


cross-modal learning / zero-shot categorisation



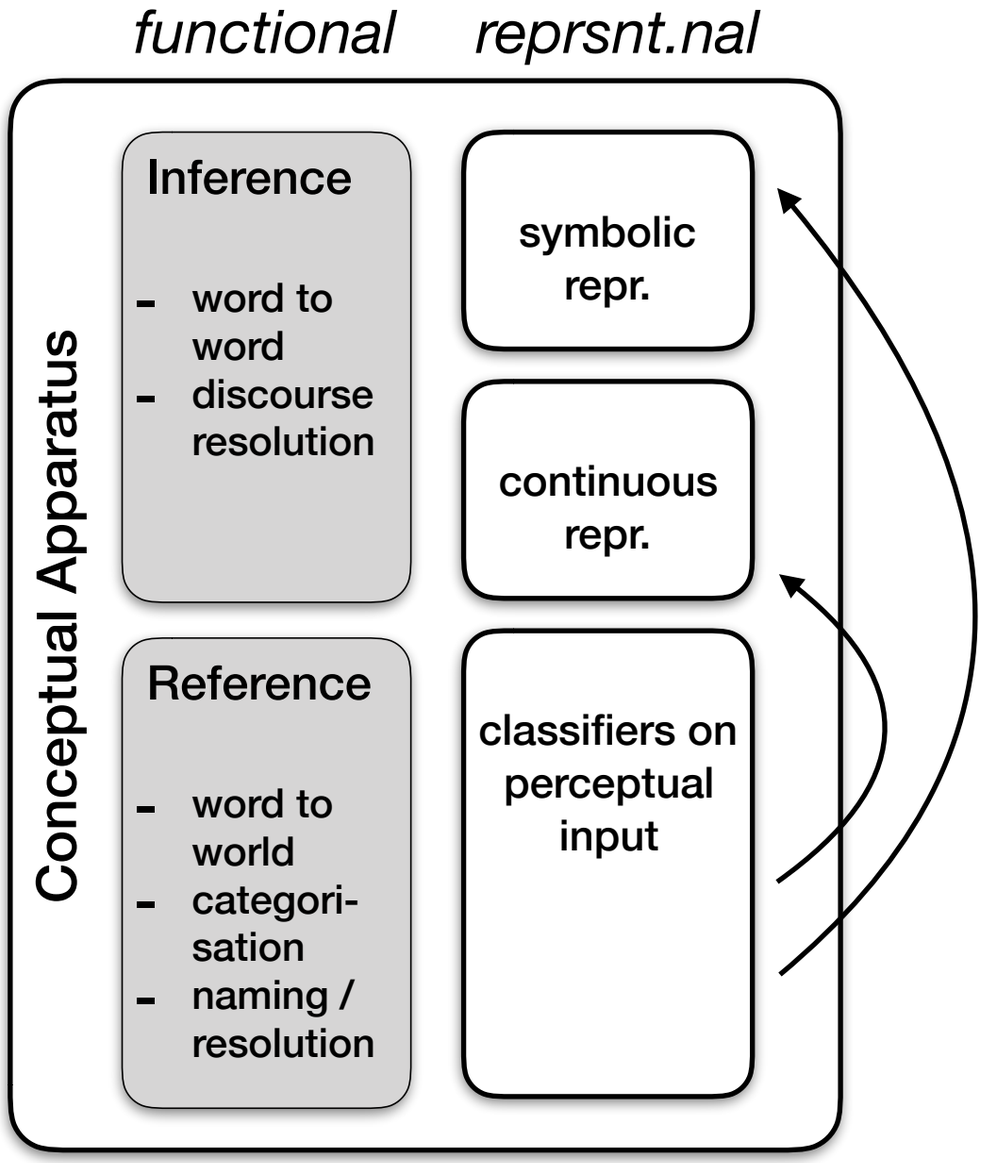
Situation:
Need to interpret RefExp (in vis. context) containing unknown word.

Approach:
Use *Principle of Contrast* (Clark 1987), assume referent is the object you can’t name.



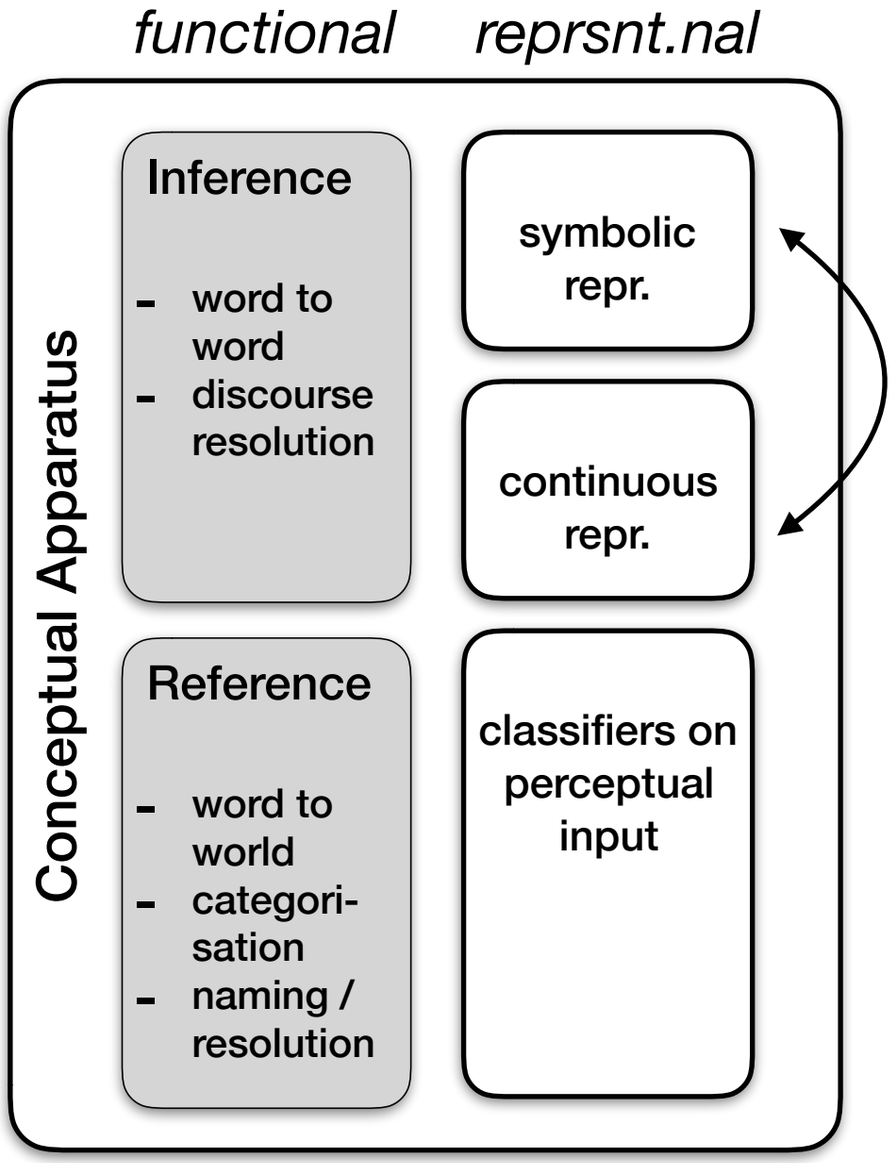
Situation:
Need to interpret RefExp (in vis. context) containing word w/o vis classifier.

Approach:
Construct classifier from neighbouring words, or from definition (Lampert *et al.* 2009).



Situation:
Need to draw inference from word for which only classifier exists.

Approach:
Learn typical attributes from instances; assume that visual similarity translates to semantic similarity and project into vector space, via neighbours.



Situation:
Encode definitions in vectors; extract definitions from distributional vectors

Approach:
Active research area.

input	background	principle	output
REX w/ unknown word	WACs for other words	only one name per category / least well recognised object	referent
REX w/ vis. unknown word	definition; WACs for words in definition	replace unknown word with definition	referent
REX w/ vis. unknown word	vector for word; WACs for neighbours	replace unknown word with neighbours	referent
REX w/ unknown word + referent	WACs for other words	concept related to visually similar ones	word vector
instance of unknown concept, identified as such	WACs, vectors	must be something in neighbourhood of best matches	name (nearest to centroid)
	vectors		theory / explicit definition
	WACs		theory / explicit definition

input	background	principle	output
REX w/ unknown word	WACs for other words	only one name per category / least well recognised object	referent

Principle of Contrast (Clark 1987): There are no synonyms.

Method: Apply all WACs (apart from “unknown” target word). Select the object where the highest scoring WAC scores lowest; it’s the one where there’s the biggest room for an unknown word.

Doesn’t work, unfortunately. The classifiers are too badly tuned to each other?

input	background	principle	output
REX w/ unknown word	WACs for other words	only one name per category / least well recognised object	referent
REX w/ vis. unknown word	definition; WACs for words in definition	replace unknown word with definition	referent

E.g., Lampert *et al.* 2009, direct attribute prediction.

classifiers for attributes

$$p(z|x) \propto \prod_{m=1}^{M} \left(\frac{p(a_m|x)}{p(a_m)} \right)^{a_m^z}$$

attributes of zero-shot class

Assemble a classifier for a category from classifiers for attributes, & knowledge of which attributes the category has.

Zero-Shot Learning with Feature Norms

behavior	eats, walks, climbs, swims, runs
diet	drinks_water, eats_anything
shape_size	is_tall, is_large
anatomy	has_mouth, has_head, has_nose, has_tail, has_claws, has_jaws, has_neck, has_snout, has_feet, has_tongue
color_patterns	is_black, is_brown, is_white

botany	has_skin, has_seeds, has_stem, has_leaves, has_pulp
color_patterns	purple, white, green, has_green_top
shape_size	is_oval, is_long
texture_material	is_shiny

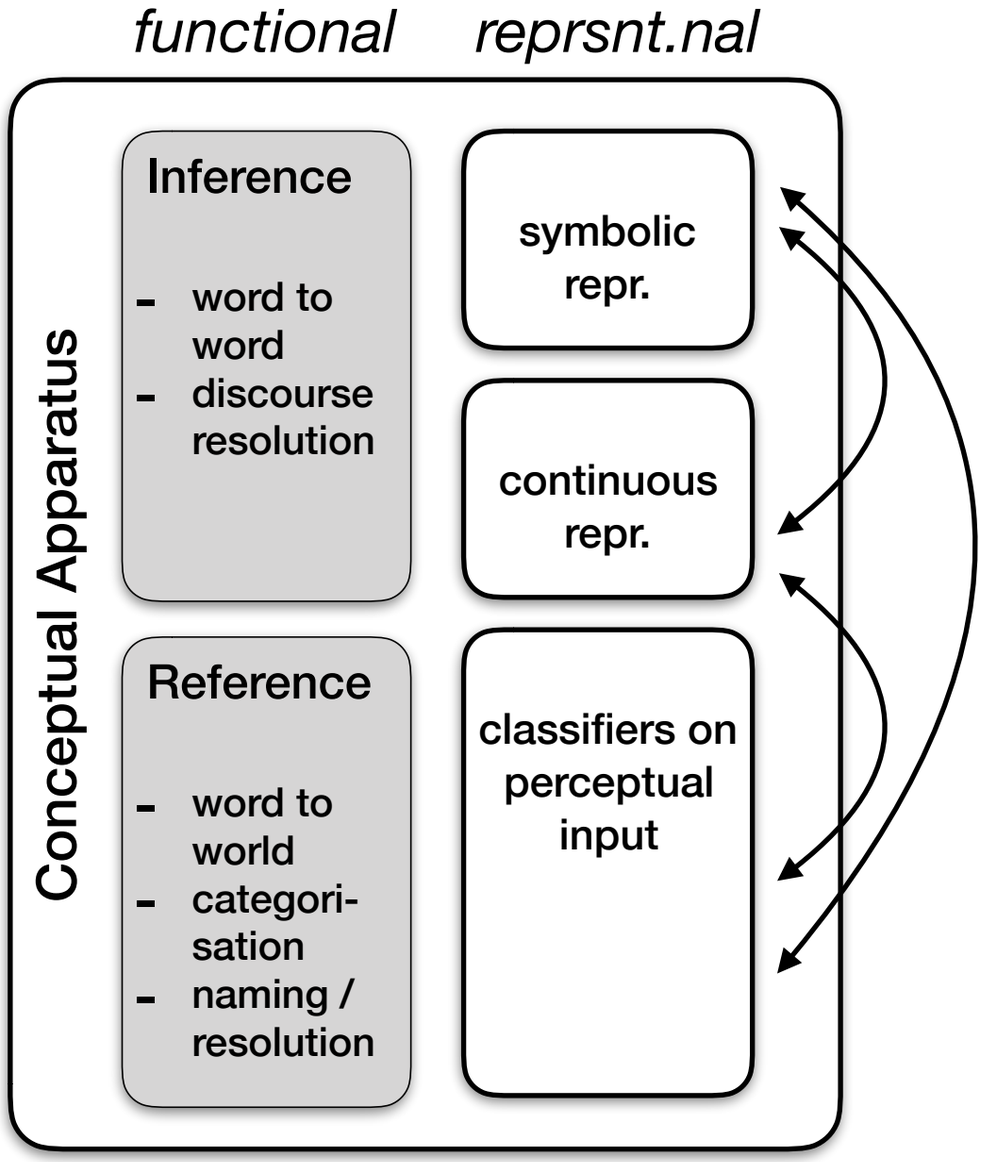
behavior	rolls
parts	has_step_through_frame, has_fork, has_2_wheels, has_chain, has_pedals has_gears, has_handlebar, has_bell, has_breaks has_seat, has_spokes
texture_material	made_of_metal
color_patterns	different_colors, is_black, is_red, is_grey, is_silver

(Silberer, Ferrari & Lapata, 2013), using feature norms of (McRae *et al.* 2005)

114 out of 509 concepts in vocab
instances for 340 of 637 attributes

Acc. on 20 test classes:
43.2%

input	background	principle	output
REX w/ unknown word	WACs for other words	only one name per category / least well recognised object	referent
REX w/ vis. unknown word	definition; WACs for words in definition	replace unknown word with definition	referent
REX w/ vis. unknown word	vector for word; WACs for neighbours	replace unknown word with neighbours	referent
REX w/ unknown word + referent	WACs for other words	concept related to visually similar ones	word vector
instance of unknown concept, identified as such	WACs, vectors	must be something in neighbourhood of best matches	name (nearest to centroid)
	vectors		theory / explicit definition
	WACs		theory / explicit definition


functional
reprsnt.nal
Conceptual Apparatus
Inference
- word to word
- discourse resolution
Reference
- word to world
- categori-sation
- naming / resolution
symbolic repr.
continuous repr.
classifiers on perceptual input

functional *reprsnt.nal* *composition* *coordination*

Conceptual Apparatus

Inference

- word to word
- discourse resolution

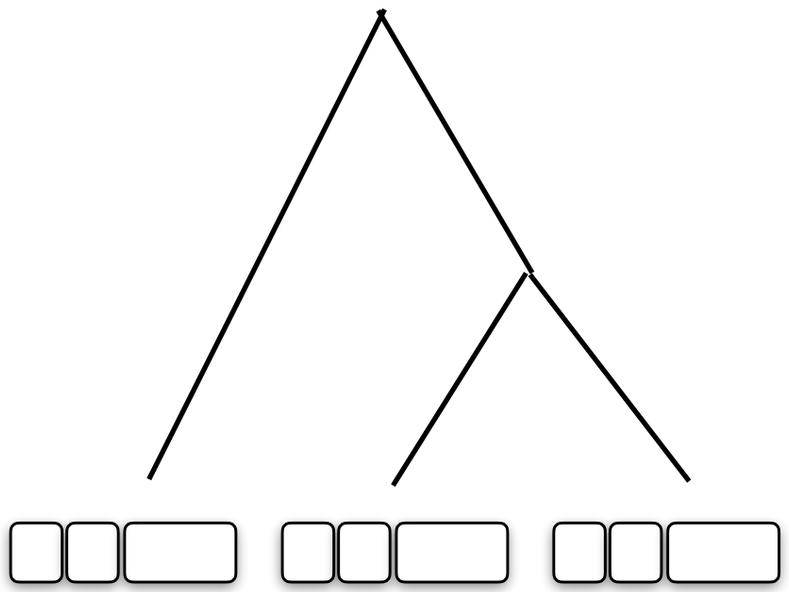
Reference

- word to world
- categori-sation
- naming / resolution

symbolic repr.

continuous repr.

classifiers on perceptual input



representations for larger expressions

- “the silver wrench on the left”… so far, treated as bag of words / predicates
- will that work for larger constructions?
- “the woman behind the laptop computer”
- “a man standing next to a walking giraffe” / “a man walking next to a standing giraffe”

representations for larger expressions

- *compositional* representations, composed out of parts

- where do we get the parts from? (and their parts..)

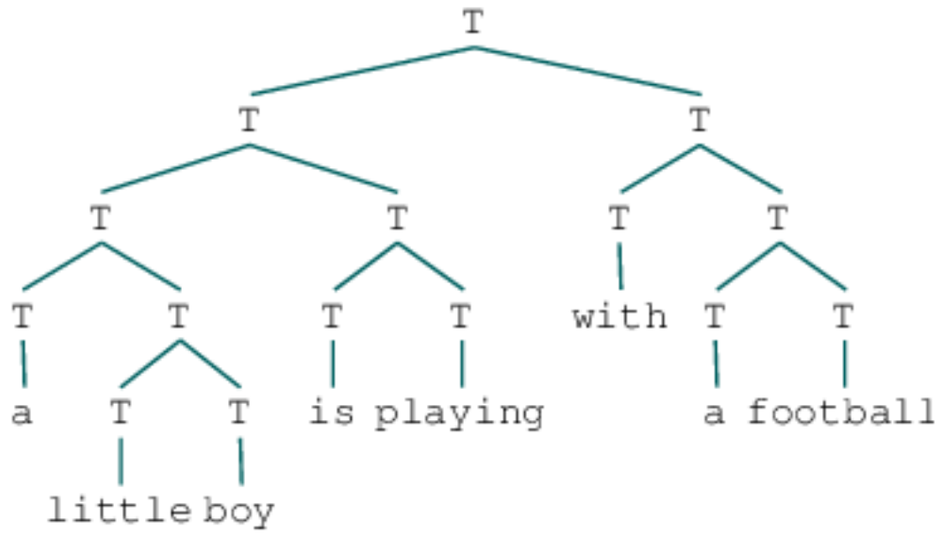
inducing structure

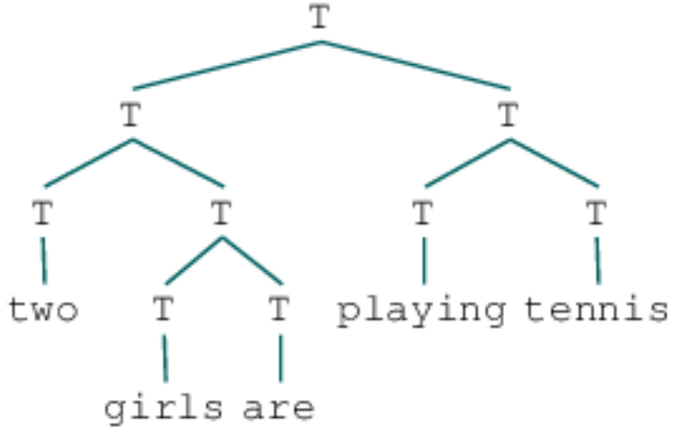
- can we learn this from visual contexts?
- idea: the contexts in which “meaningful sub-parts” occur are more mutually similar (visually) than those of “non-meaningful sub-parts”.
- tested with captions on mscoco

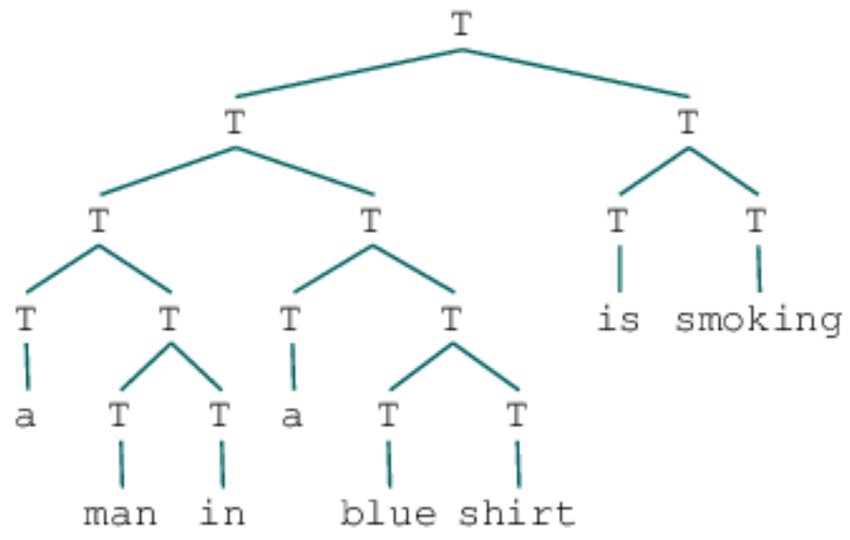
```
(('on', 'a', 'stove'), 0.117957666951
(('men', 'are', 'cooking'), 0.10513341
(('in', 'a', 'kitchen'), 0.09961850608
(('cooking', 'a', 'meal'), 0.099154587
(('preparing', 'food', '.'), 0.0985616
(('cooking', 'food', 'on'), 0.09761229
(('holds', 'a', 'guitar'), 0.091528381
(('on', 'a', 'cake'), 0.090694738143 10
(('park', 'bench', 'with'), 0.08527560
(('a', 'meal', '.'), 0.085138958440741
(('a', 'stove', '.'), 0.08407311167122
(('the', 'guitar', 'and'), 0.082618504
(('a', 'cake', '.'), 0.081963140581116
(('a', 'living', 'room'), 0.0811017693
(('in', 'a', 'living'), 0.081060728493
(('living', 'room', '.'), 0.0805577166
```

inducing structure

- use this to score all possible binary trees (bigrams & trigrams)

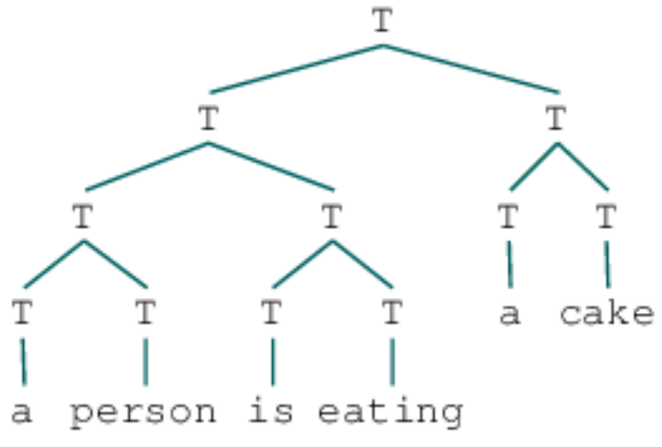




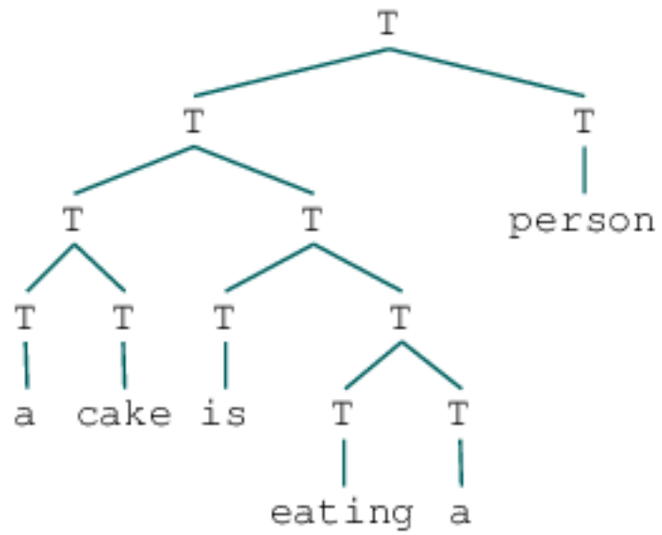


inducing structure

- use this to score all possible binary trees (bigrams & trigrams)



- how to evaluate?



structure

- for now, lets assume we have a parser that gives us a structure for an input expression

- how can we use that to guide construction of representations of expressions beyond single words?

- we will use the same syntactic structure to guide computation of reference and computation of continuous representation for phrase

- we'll start with continuous representations

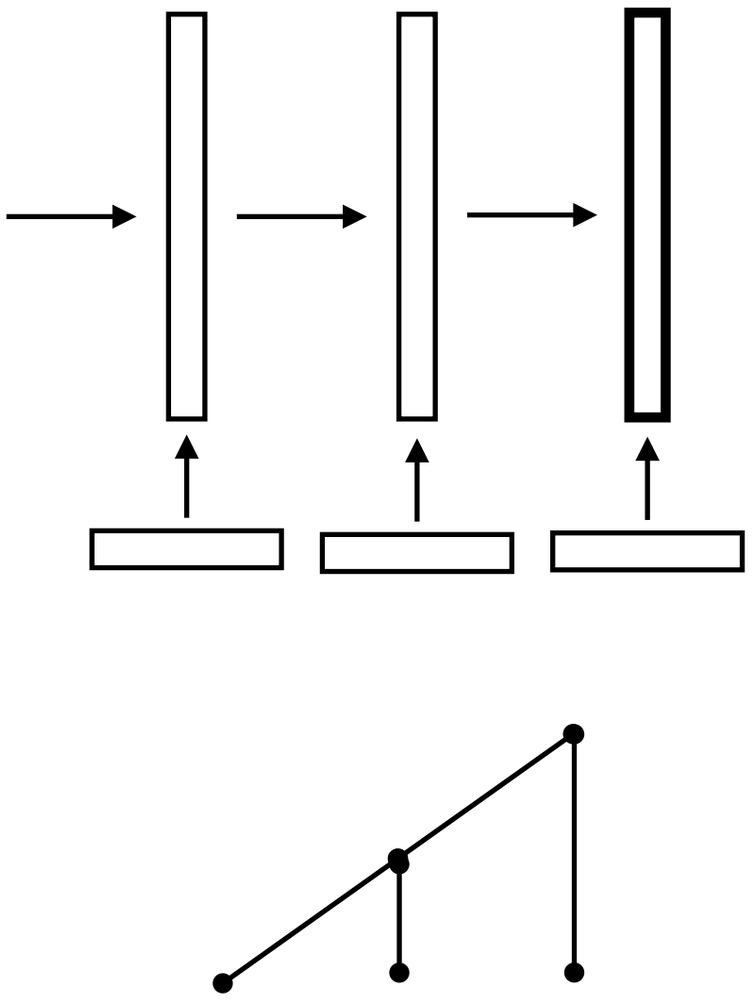
structure

- simplest way to get a continuous representation for phrase: simply add up the word representations.

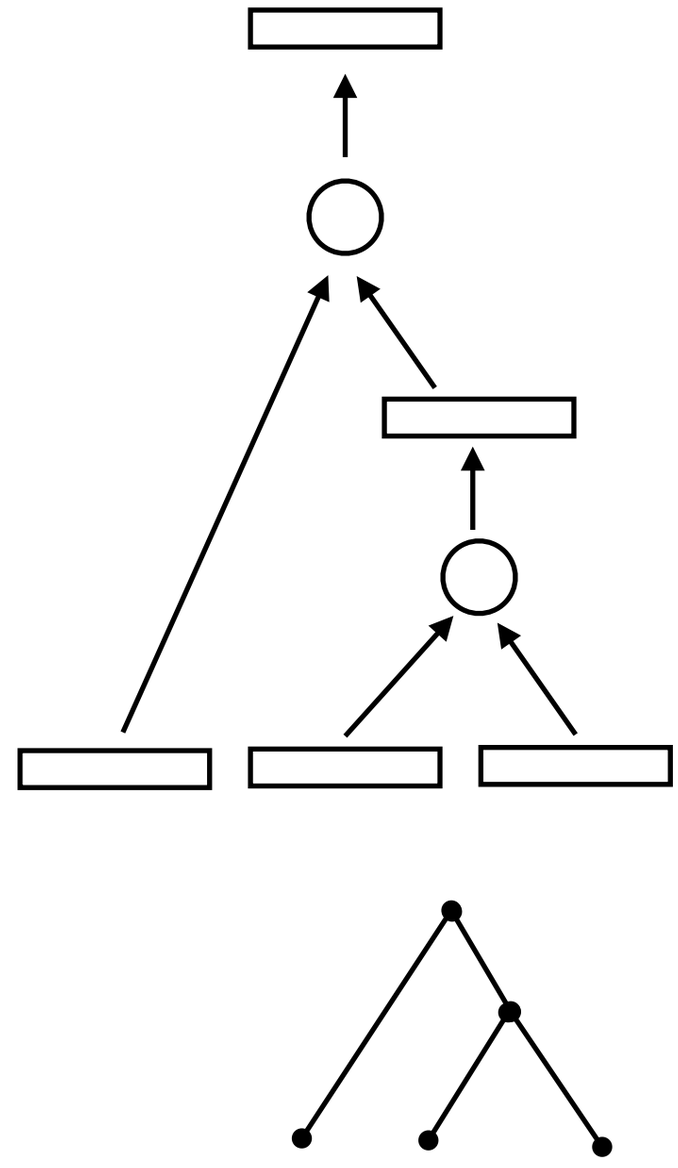
- (“simple, but tough-to-beat”, [Arora *et al.* 2017])

- but that doesn’t respect structure, still bag-of-words (“a man standing next to a walking giraffe” / “a man walking next to a standing giraffe”)

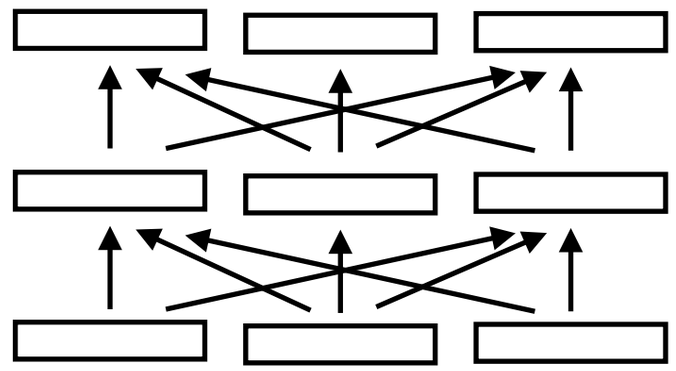
using neural networks to construct phrase representations



(Elmann 1990,
Hochreiter & Schmidhuber 1997)



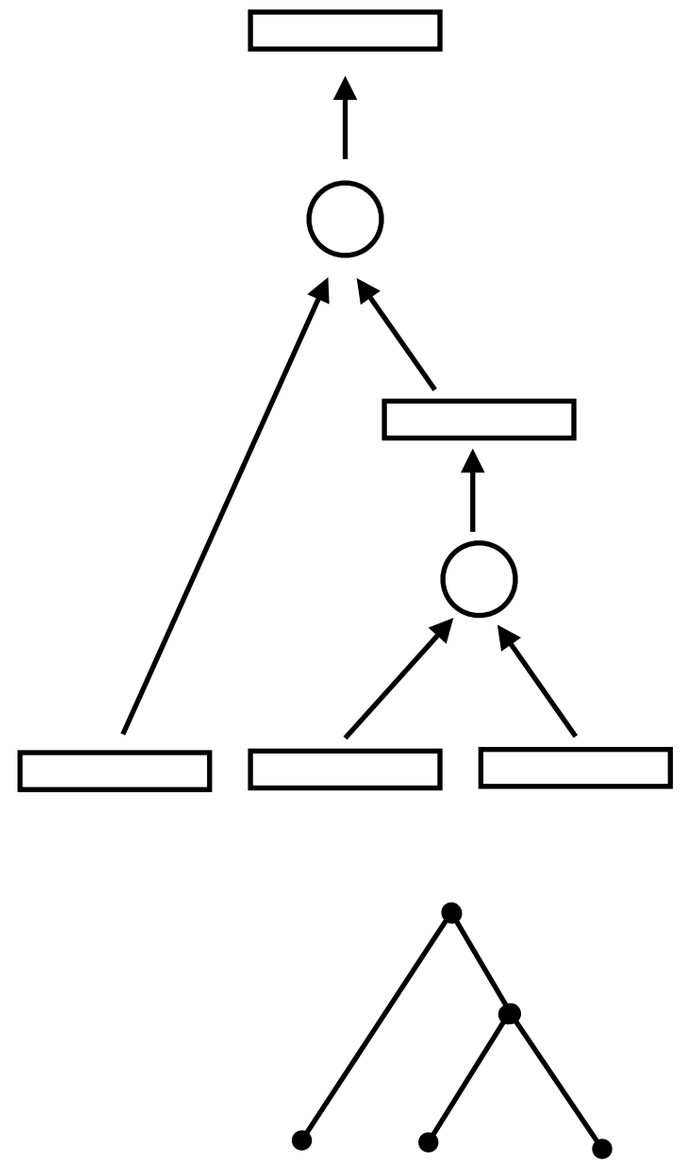
(Socher *et al.* 2011)



(Vaswani *et al.* 2017)

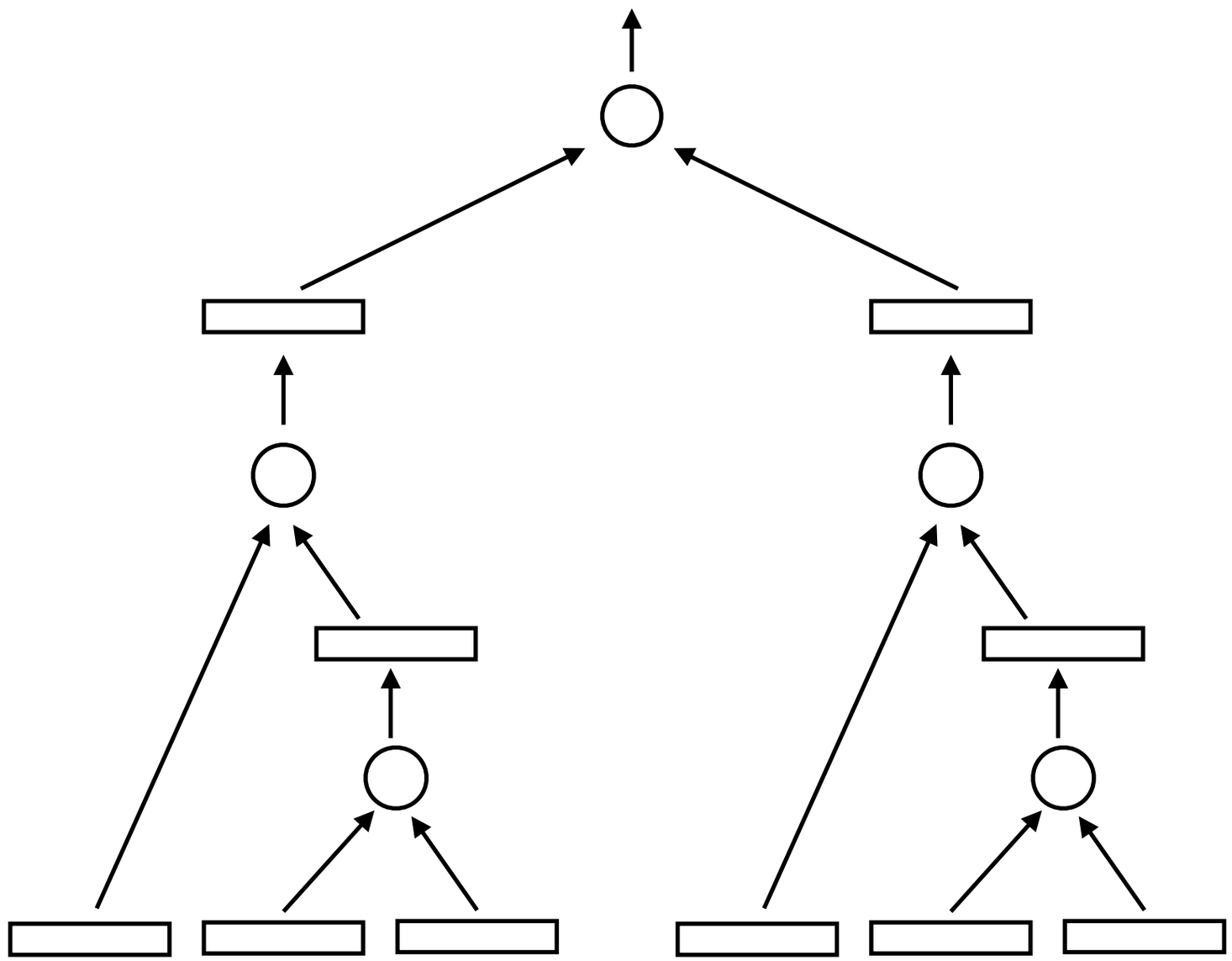
using neural networks to construct phrase representations

We have used TreeLSTMs, as implemented in SPINN (Bowman *et al.* 2016).



(Socher *et al.* 2011)

using continuous phrase representations for inference tasks



(Bowman *et al.* 2015)

task 1: rex / rex

- premise: a referring expression
- hypothesis: another referring expression, from same image
- question: can the hypothesis be used in reference to the same object as the premise?
- data: validation sections of corpora (to train model), test section

```
neutral: person on right reaching out || left guy
neutral: person on right reaching out || man white shirt back
neutral: person right || man white shirt back
entailment: white shirt || man white shirt back
neutral: white shirt || woman in purple
entailment: man on left || old guy
entailment: man on left || left guy
entailment: old guy || left guy
neutral: man on left || man white shirt back
neutral: man on left || woman in purple
neutral: old guy || white shirt
entailment: person above laptop || jeans above silver laptop
entailment: person above laptop || jeans
entailment: jeans above silver laptop || jeans
neutral: person above laptop || laptop with green keyboard
neutral: person above laptop || jeans on right
neutral: checkered tablecloth || olive on pizza in front
neutral: checkered table cloth || right pizza
entailment: serving surface || table that the food is on
neutral: left bear || right bear
```

task 2: cap / object

- premise: a caption
- hypothesis: the name of an object
- question: is the object likely to be found in the scene described by the caption?
- data: visgen

```
entailment: A kitchen with a pets food dishes on the floor. || cabinets
entailment: A kitchen with a pets food dishes on the floor. || black
neutral: A kitchen with a pets food dishes on the floor. || trees
neutral: A kitchen with a pets food dishes on the floor. || cloud
entailment: A very clean kitchen with some animal food on the ground. ||
dish soap
entailment: A very clean kitchen with some animal food on the ground. ||
rack
neutral: A very clean kitchen with some animal food on the ground. ||
kite
neutral: A very clean kitchen with some animal food on the ground. ||
pole
entailment: A man on top of a snowy mountain with skis. || hat
entailment: A man on top of a snowy mountain with skis. || goggles
neutral: A man on top of a snowy mountain with skis. || small airplane
neutral: A man on top of a snowy mountain with skis. || zebra
entailment: A group of skiers stand atop a snowy peak. || ski
entailment: A group of skiers stand atop a snowy peak. || backpack
neutral: A group of skiers stand atop a snowy peak. || basket
neutral: A group of skiers stand atop a snowy peak. || train tracks
```

task 3: cap / cap

- premise: a caption
- hypothesis: another caption
- question: does the hypothesis describe the same scene?
- data: visgen; selecting distractors according to visual similarity

neutral: A man cutting food with a pair of blue scissors. || Three cats are on the table watching the tv together.

neutral: A man sitting with his back up against a tree. || Three cats sitting on a counter watching television.

entailment: Two young children on a ramp with their skate boards. || Young boys in raincoats skate boarding on a half pipe

entailment: a person on a skate board rides up a ramp || A kid dressed in black is on his skateboard on a skateboard ramp.

neutral: Two young children on a ramp with their skate boards. || Two kids are on skis at the top of a slope.

neutral: Two young children on a ramp with their skate boards. || Young children on snowy area of alpine region.

neutral: People are sitting in the dark using computers. || a big t.v. that has a cat sitting next to it

neutral: The people are using their computers in the dark. || Two laptop computers sitting on top of a wooden table.

entailment: A bowl of red soup and a sandwich cut in half. || a close up of a sandwich on a plate near a spoon

neutral: A table topped with a couple of sandwiches and a bowl of soup. || a plate of food near a drink on a table

results

	task 1 (rex/rex)	task 2 (cap/obj)	task 3 (cap/cap)
baseline (overlap)	69%	57%	53%
TreeLSTM	82%	76%	79%

Conclusion: Learned representations capture gist of described situation?

Still to do: test with actual co-reference resolution task, rhetorical relation prediction task.

functional *reprsnt.nal* *composition* *coordination*

Conceptual Apparatus

Inference

- word to word
- discourse resolution

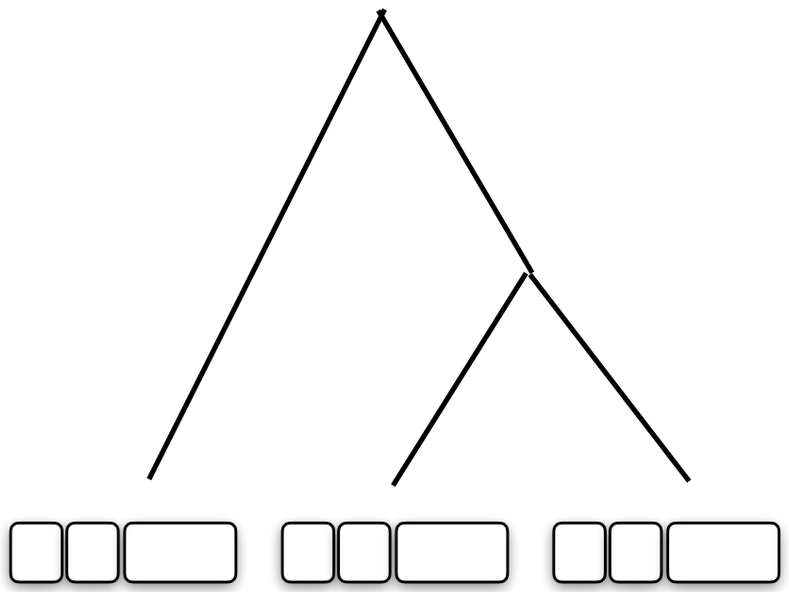
Reference

- word to world
- categori-sation
- naming / resolution

symbolic repr.

continuous repr.

classifiers on perceptual input



functional *reprsnt.nal* *composition* *coordination*

Conceptual Apparatus

Inference

- word to word
- discourse resolution

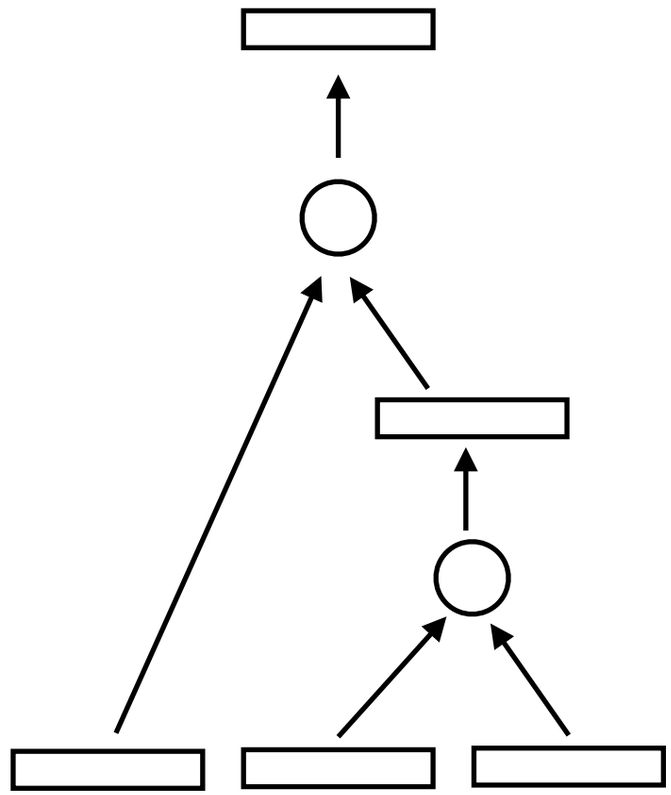
Reference

- word to world
- categori-sation
- naming / resolution

symbolic repr.

continuous repr.

classifiers on perceptual input



syntactic structure

- where did the parses come from?
- from a large scale, wide-coverage HPSG grammar!
- … which also produces fine-grained semantic representations

English Resource Grammar (Flickinger 2000, 2011)
erg.delph-in.net

DELPH-IN

- Under continuous development since 1993

- Broad-coverage: 85-95% on varied domains: newspaper text, Wikipedia, bio-medial research literature (Flickinger et al 2010, 2012; Adolphs et al 2008)
  - Robust processing techniques enable 100% coverage

- Output: derivation trees paired with meaning representations in the Minimal Recursion Semantics framework---English Resource Semantics (ERS)

  - Emerging documentation at moin.delph-in.net/ErgSemantics

(slide from presentation by Emily Bender, 2019)

parsing the corpus

- using the “Answer Constraint Engine” (ACE; Packard 2018; http://sweaglesw.org/linguistics/ace/), I parsed (large parts of) the referring expression and caption corpora

- ~90% of expressions had at least one analysis

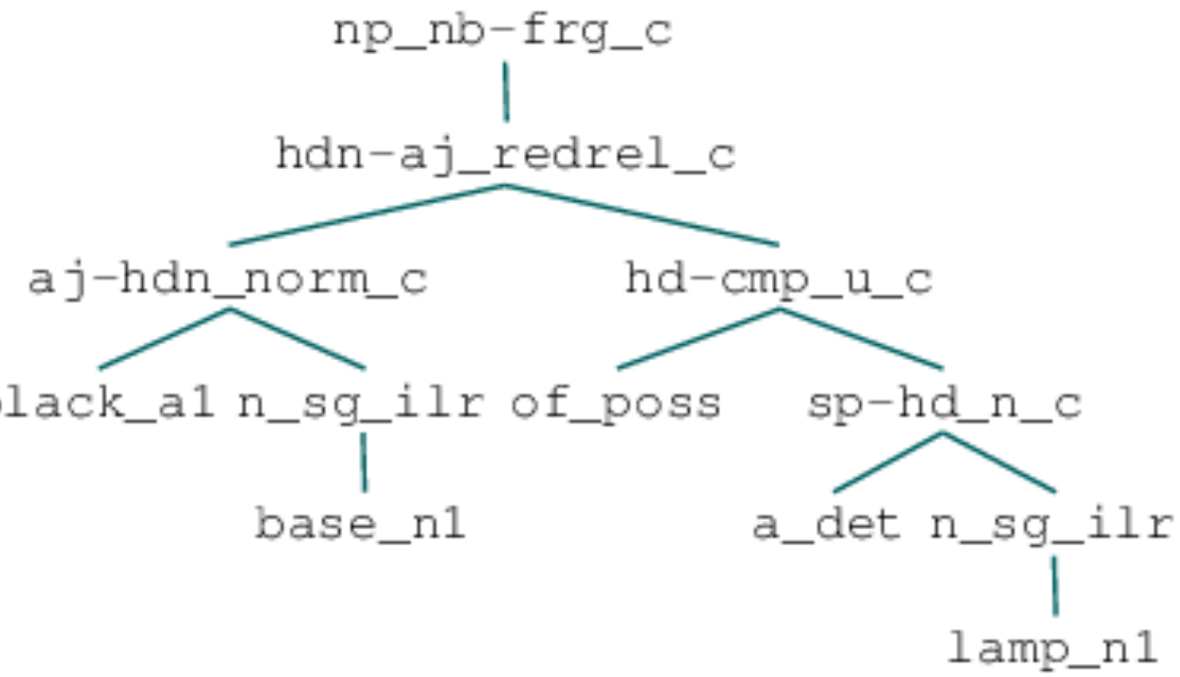
- no hard numbers; intuitionistically, about 70% of semantic representations somewhat sensible

- frequent problems with reduced relative clauses, elided determiners, nouns interpreted as verbs, weird compounds: “skateboard on the floor”, “two girls stand next to a dog”, …

- but still gets NPs and relations right often (enough)

examples

Black base of a lamp

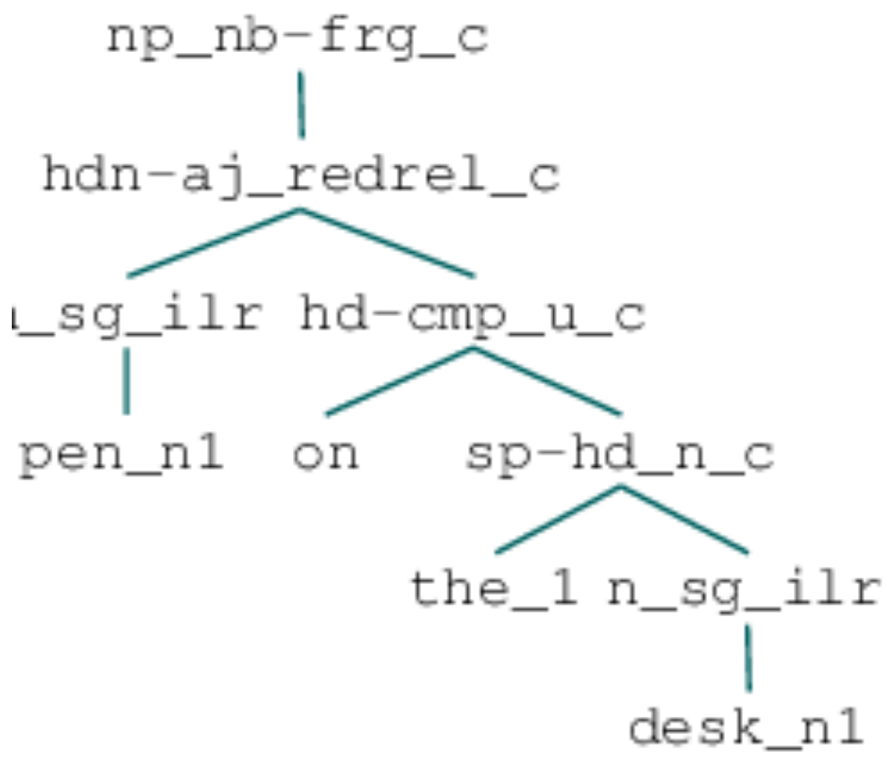



```
h1 [('e2', 'unknown', None, {'ARG0': 'e2', 'ARG': 'x4'})]
h5 [('x4', 'udef_q', 'q', {'ARG0': 'x4', 'RSTR': 'h6', 'BODY': 'h7'})]
h8 [('e9', '_black_a_1', 'a', {'ARG0': 'e9', 'ARG1': 'x4'}), ('x4',
'_base_n_1', 'n', {'ARG0': 'x4'}), ('e10', '_of_p', 'p', {'ARG0': 'e10',
'ARG1': 'x4', 'ARG2': 'x11'})]
h12 [('x11', '_a_q', 'q', {'ARG0': 'x11', 'RSTR': 'h13', 'BODY': 'h14'})]
h15 [('x11', '_lamp_n_1', 'n', {'ARG0': 'x11'})]

{'x4': ('base', ('1072501', 'base.n.01')),
 'x11': ('lamp', ('1072502', 'lamp.n.01'))}
```

examples

pen on the desk




```
h1 [('e2', 'unknown', None, {'ARG0': 'e2', 'ARG': 'x4'})]
h5 [('x4', 'udef_q', 'q', {'ARG0': 'x4', 'RSTR': 'h6', 'BODY': 'h7'})]
h8 [('x4', '_pen_n_1', 'n', {'ARG0': 'x4'}), ('e9', '_on_p_loc', 'p', {'ARG0': 'e9',
'ARG1': 'x4', 'ARG2': 'x10'})]
h11 [('x10', '_the_q', 'q', {'ARG0': 'x10', 'RSTR': 'h12', 'BODY': 'h13'})]
h14 [('x10', '_desk_n_1', 'n', {'ARG0': 'x10'})]
```

composing reference computation graphs

```
h1 [('e2', 'unknown', None, {'ARG0': 'e2', 'ARG': 'x4'})]
h5 [('x4', 'udef_q', 'q', {'ARG0': 'x4', 'RSTR': 'h6', 'BODY': 'h7'})]
h8 [('e9', '_black_a_1', 'a', {'ARG0': 'e9', 'ARG1': 'x4'}), ('x4',
'_base_n_1', 'n', {'ARG0': 'x4'}), ('e10', '_of_p', 'p', {'ARG0': 'e10',
'ARG1': 'x4', 'ARG2': 'x11'})]
h12 [('x11', '_a_q', 'q', {'ARG0': 'x11', 'RSTR': 'h13', 'BODY': 'h14'})]
h15 [('x11', '_lamp_n_1', 'n', {'ARG0': 'x11'})]
```

```
{'x4': ('base', ('1072501', 'base.n.01')),
 'x11': ('lamp', ('1072502', 'lamp.n.01'))}
```

- we train classifiers for unary predicates (e.g., `_black_a_1`),
- … and for binary predicates (`_of_p`, `_on_loc_e`), using the positional features only [i.e., only spatial interpretation of relations for now]

composing reference computation graphs

```
h1 [('e2', 'unknown', None, {'ARG0': 'e2', 'ARG': 'x4'})]
h5 [('x4', 'udef_q', 'q', {'ARG0': 'x4', 'RSTR': 'h6', 'BODY': 'h7'})]
h8 [('e9', '_black_a_1', 'a', {'ARG0': 'e9', 'ARG1': 'x4'}), ('x4',
'_base_n_1', 'n', {'ARG0': 'x4'}), ('e10', '_of_p', 'p', {'ARG0': 'e10',
'ARG1': 'x4', 'ARG2': 'x11'})]
h12 [('x11', '_a_q', 'q', {'ARG0': 'x11', 'RSTR': 'h13', 'BODY': 'h14'})]
h15 [('x11', '_lamp_n_1', 'n', {'ARG0': 'x11'})]
```

```
{'x4': ('base', ('1072501', 'base.n.01')),
 'x11': ('lamp', ('1072502', 'lamp.n.01'))}
```

Application

- intersection for predicates of same variable (as before)
- relations evaluated on all pairs of objects, add weight to interpretations of NPs (e.g. “dog on chair”, the best dog-like object that is also most “on chair”)
- quantifiers have hard-coded function, e.g. “the” as argmax
- semantic types: NPs to entities, Ss to truth values

composing reference computation graphs

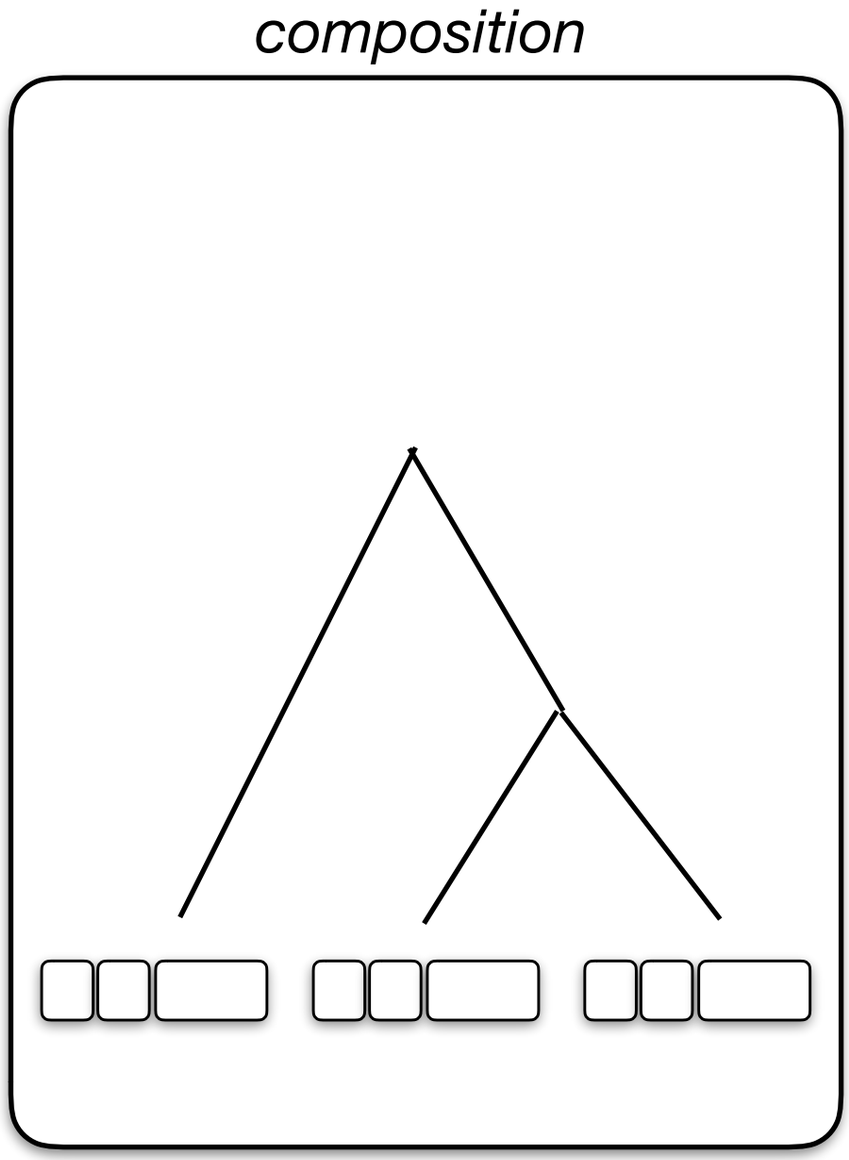
```
h1 [('e2', 'unknown', None, {'ARG0': 'e2', 'ARG': 'x4'})]
h5 [('x4', 'udef_q', 'q', {'ARG0': 'x4', 'RSTR': 'h6', 'BODY': 'h7'})]
h8 [('e9', '_black_a_1', 'a', {'ARG0': 'e9', 'ARG1': 'x4'}), ('x4',
'_base_n_1', 'n', {'ARG0': 'x4'}), ('e10', '_of_p', 'p', {'ARG0': 'e10',
'ARG1': 'x4', 'ARG2': 'x11'})]
h12 [('x11', '_a_q', 'q', {'ARG0': 'x11', 'RSTR': 'h13', 'BODY': 'h14'})]
h15 [('x11', '_lamp_n_1', 'n', {'ARG0': 'x11'})]
```

```
{'x4': ('base', ('1072501', 'base.n.01')),
 'x11': ('lamp', ('1072502', 'lamp.n.01'))}
```

- promissory note in (Schlangen *et al.* ACL 2016)
- neural modules, e.g. (Andreas *et al.* 2016). But they use ad-hoc functions (`locate`, `filter`, etc.); we try to re-use normal formal semantics interpretation rules

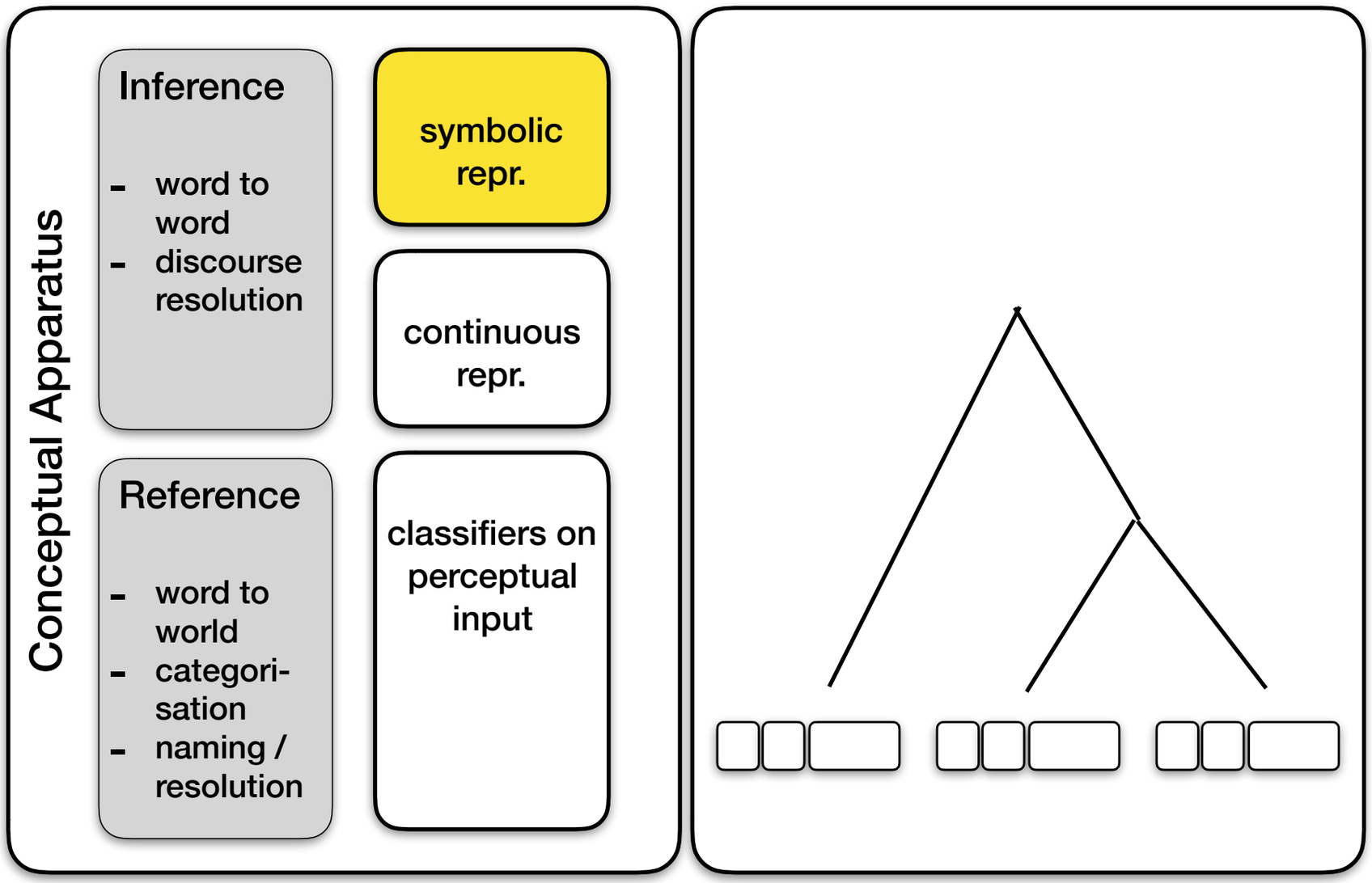
results

- visual genome region descriptions, treated as referring expressions picking out single object in scene:
  - @1: 25% accuracy (@5: 61%).. random baseline: 3%
    - w/o relations, only NPs: @1: 24% (@5: 58%)
- treated as statements (“there is a”):
  - learn threshold on validation set, false statements = region descriptions paired with wrong image
  - 67% accuracy in judging correctness of statement



Classifiers: Use logical form as computation graph for computing reference.

Vectors: Use syntactic structure to guide recursive composition into phrase representations.



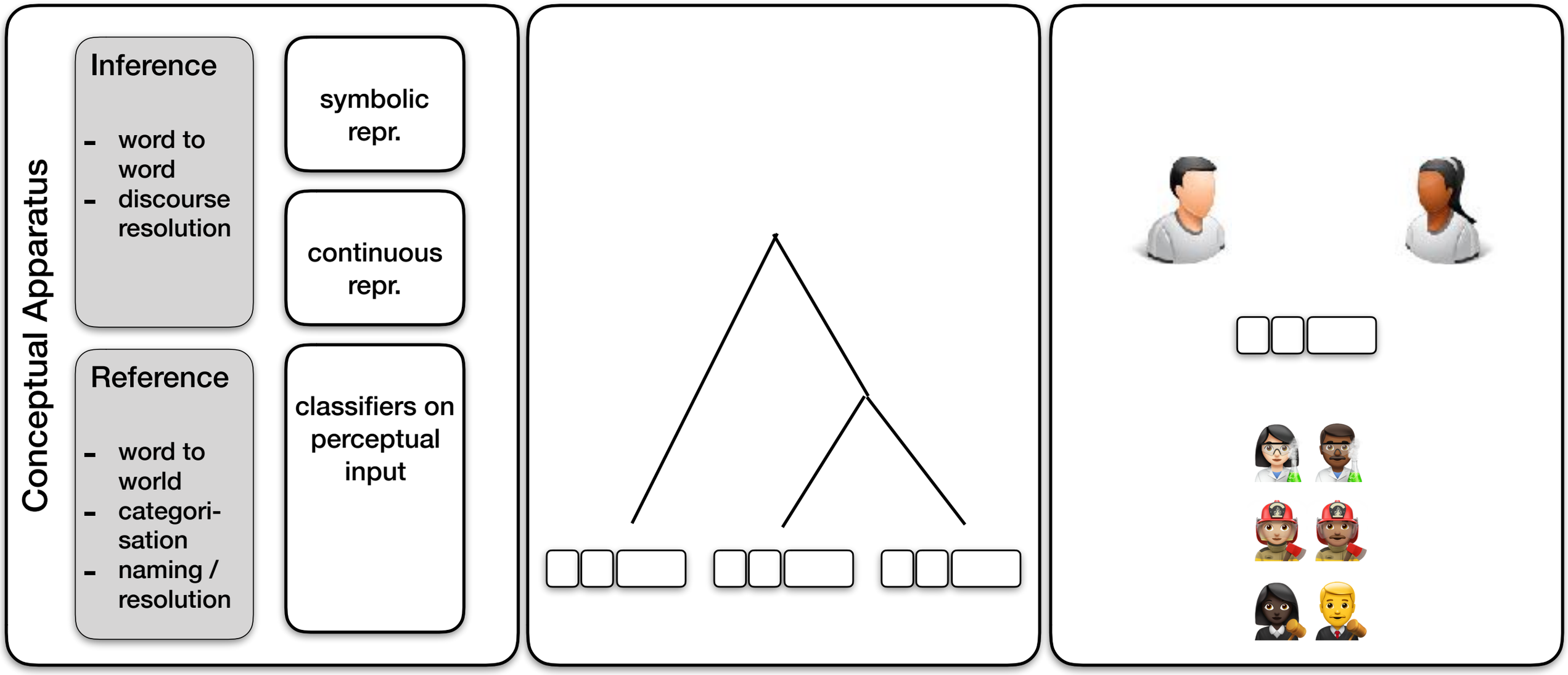
Classifiers: Use logical form as computation graph for computing reference.

Vectors: Use syntactic structure to guide recursive composition into phrase representations.

We also still have the logical form around to do “classical” inference. And the definitions to do recognition…

Isn’t that a bit much?

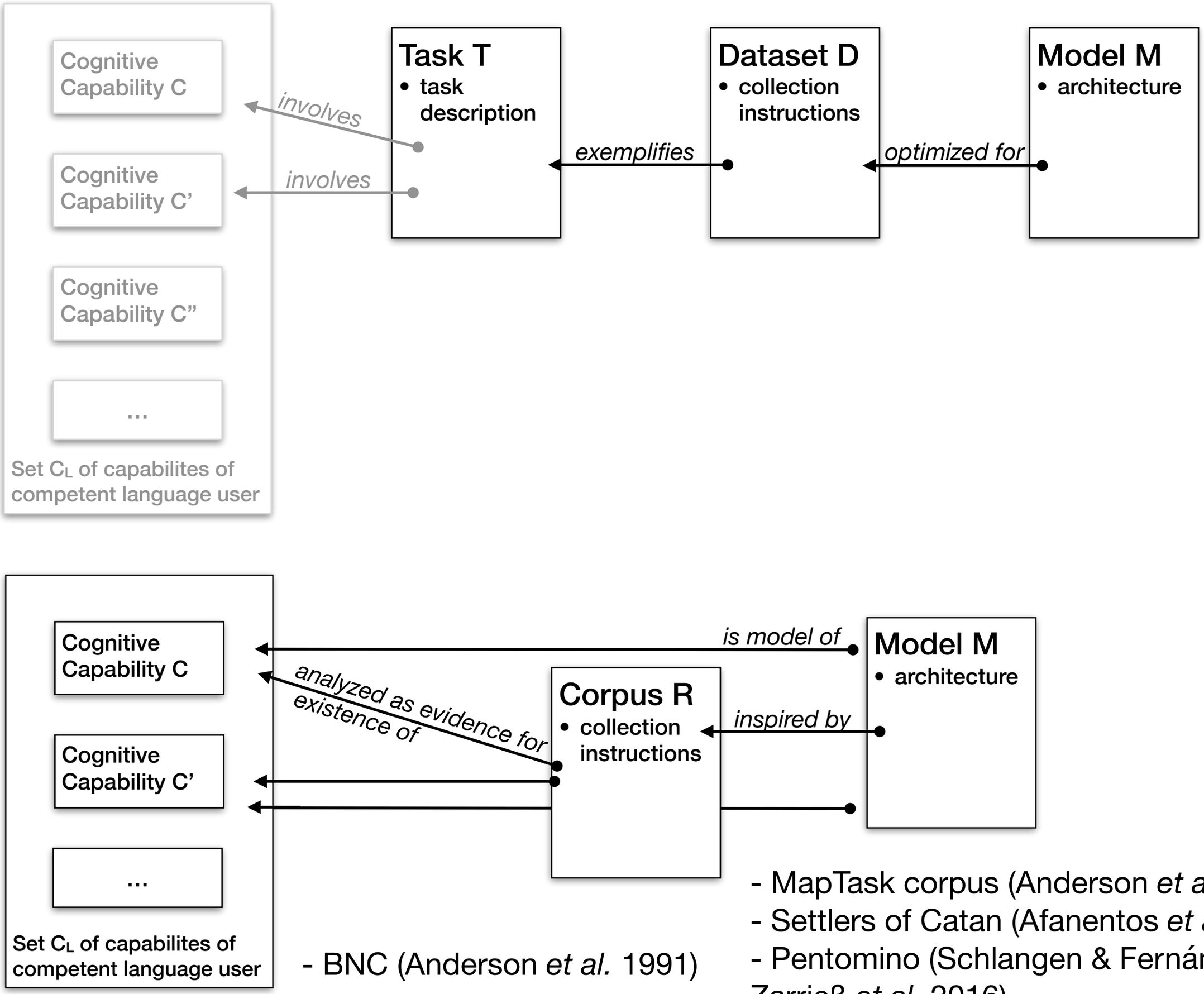
Idea is that *“slow” inference* can kick in when needed for *justification* and *deliberation.* Symbolic inference and symbolic definitions are things that we can talk about.



We also still have the logical form around to do “classical” inference. And the definitions to do recognition…

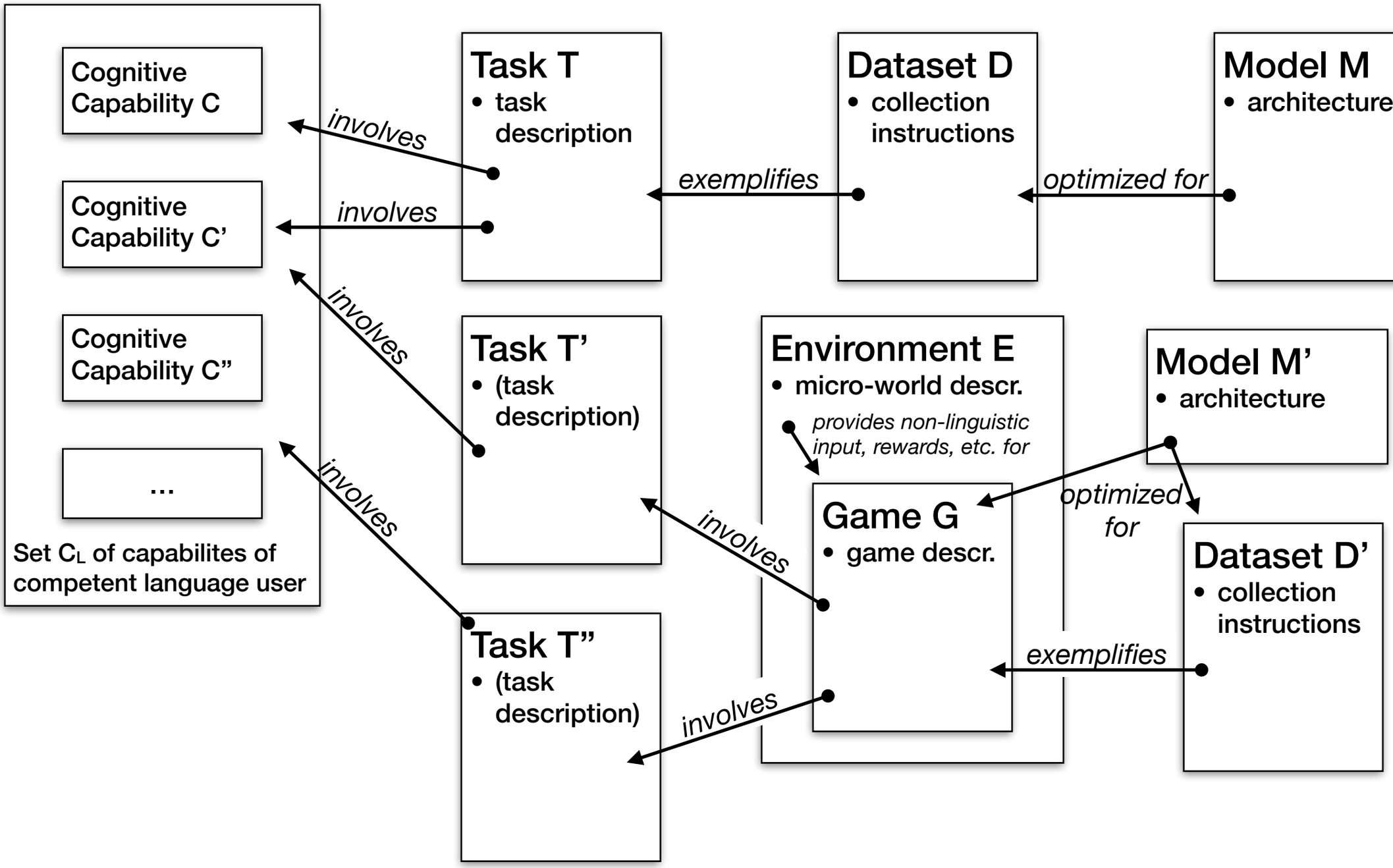
Isn’t that a bit much?

Idea is that *“slow” inference* can kick in when needed for *justification* and *deliberation.* Symbolic inference and symbolic definitions are things that we can talk about.



- BNC (Anderson *et al.* 1991)

- MapTask corpus (Anderson *et al.* 1991)
- Settlers of Catan (Afanentos *et al.* 2012)
- Pentomino (Schlangen & Fernández 2008, Zarrieß *et al.* 2016)



"Language Tasks and Language Games: On Methodology in Current NLP Research", Schlangen 2019, ArXiv
"MeetUp!" Schlangen *et al.* semdial 2019

Related Work: Visual Dialogue

a grandfather clock fits into the corner of an ornately decorated living room

is the grandfather looked antique?	yes
is it tall?	yes
is it painted?	the clock is brown wood
what time does it tell?	the photo doesn't really show that
are there couch and sofa?	yes
is it antique too?	no
is the room clean?	yes
is the whole living modern?	it seems to be
what other furniture are there?	coffee table
are there people?	no

(Das *et al.*, CVPR 2017)

Related Work: Guess What

```
['Is it food?' 'No']
['Is it a utensil?' 'Yes']
['Is it the wooden spoon?' 'No']
['Is it the knife?' 'Yes']
```

(de Fries *et al.*, CVPR 2017)

Related Work:
Image Grounded Conversations

A: I want my coffin to be moved like this!
B: You wouldn't mind holding up traffic for miles?
A: No, because it will be that important when I die!
B: Are you going to have an enormous headstone on your grave also?
A: Yes, naturally.
B: Well it's the last thing you'll do so I guess you should make it count!

(Mostafazadeh *et al.,* IJCNLP 2017)

Related Work:
Vision-and-Language Navigation

Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

(Anderson *et al.,* CVPR 2018)

The Missing Ingredient: A Joint Purpose

An *Agreement Game* is a dialogue game with two regular participants, $\mathcal{P} = \{P_1, P_2\}$, and a disinterested third participant, N (for *Nature*). N poses a question Q to the players $\mathcal{P}$, and provides them with information I required to answer the question; possibly split up over the players. If I contains visual information, we call the game a *Grounded* Agreement Game.

The players can exchange messages in an unrestricted way. The game ends when one of the players explicitly proposes an answer A and the other player explicity agrees with the proposal. As the answer A will be based on a construal of I, the agreement on A is also an agreement on that construal. Optionally, a reward can be given to the players after they have provided their joint answer, tied to some measure of quality of A.

„Grounded Agreement Games: Emphasizing Conversational Grounding in Visual Dialogue Settings“
(Schlangen 2019, ArXiv)

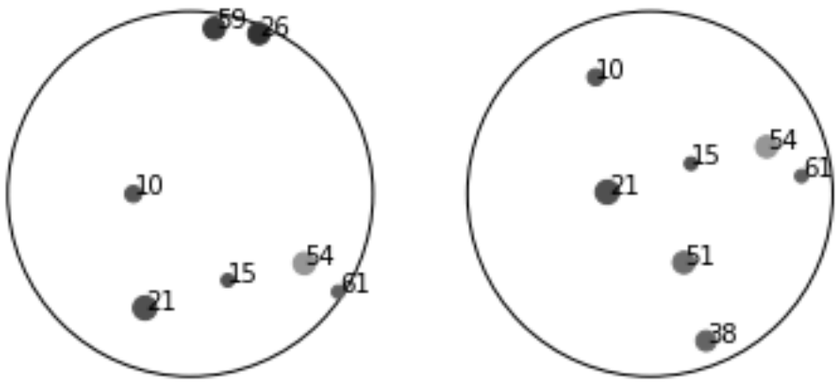
Related Work: PhotoBook

timestamp	message	speaker
00:00	H**lo!	A
00:18	Do you have a picture that has a blue bus in it that says IKEA?	A
00:31	Yes and Do you have a red truck towing a bus	B
00:34	**com** bus_truck/COCO_train2014_000000483534.jpg	B
00:34	**com** bus_truck/COCO_train2014_000000483534.jpg	A
00:35	**com** bus_truck/COCO_train2014_000000483534.jpg	A
00:44	I don't	A
00:50	**dif** bus_truck/COCO_train2014_000000349204.jpg	B
01:06	Do you have a blue truck pulling a white and yellow bus	B
01:13	I do	A

timestamp	message	speaker
03:50	Turquise bus?	B
03:54	No	A
03:58	**dif** bus_truck/COCO_train2014_000000400091.jpg	B
03:59	Guy on bike?	A
04:05	**com** bus_truck/COCO_train2014_000000557394.jpg	B
04:06	Yes	B
04:08	**com** bus_truck/COCO_train2014_000000557394.jpg	A
04:21	Red truck pulling red bus?	A
04:22	Guy on yellow/green motorcycle?	B
04:35		
04:38	**dif** bus_truck/COCO	

(Haber *et al.*, ACL 2019)

Related Work:
One in Common

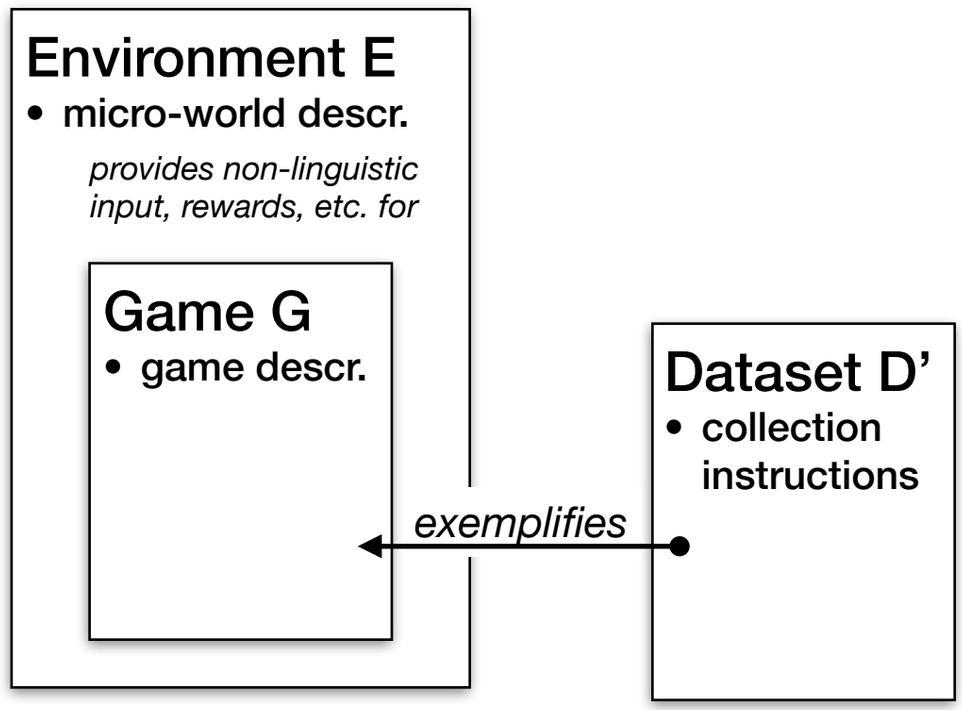


agent	data	start_time	time
0	I have two dark twins, very close to each other, and large.	0	11.99
1	I don't see those. I have one light grey one, small darker grey below it to the right?	20.05	68.93
0	I see those, I think. There is another small dark one to the lower left of the big light one?	84.12	106.3
1	Yes, further away.	118.83	121.22
0	Yes! Let's go with the large light colored one!	125.12	134.41
1	Good.	137.79	138.72
0	54	nan	139.65
1	54	nan	140.16

(Udagawa & Aizawa, AAAI 2019)

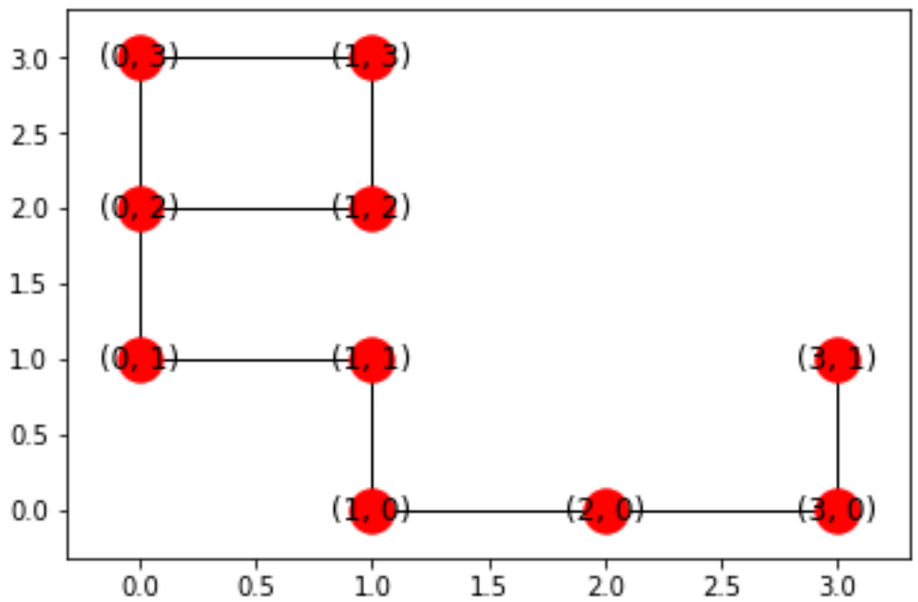
Overview

- Motivation: Progress through Datasets, Tasks, and Games
- Related Work: From One-Sided Dialogue to Grounded Agreement Games
- The MapWorld Environment
- The MeetUp-X Game
- The Dataset
- Conclusions



The MapWorld Environment

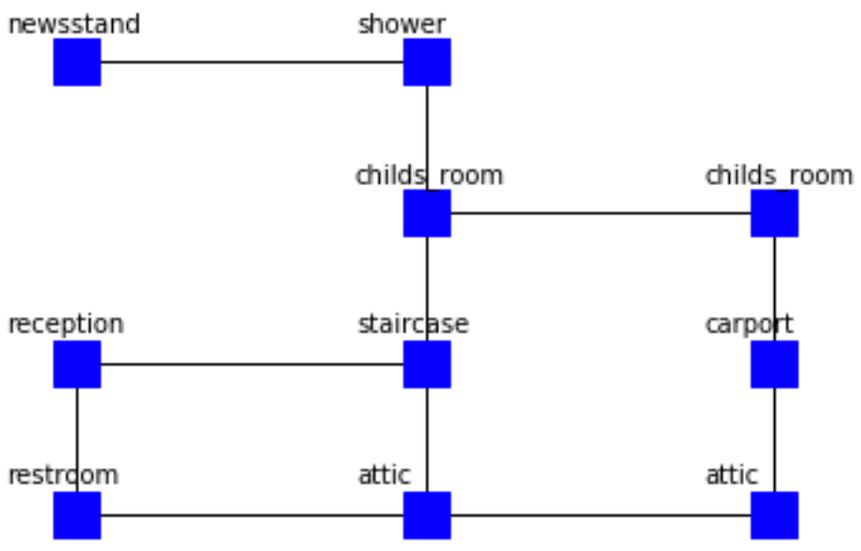
- an “abstract map” (connected subgraph of two-dimensional grid graph), created by random 2d walk



The MapWorld Environment

- an “abstract map” (connected subgraph of two-dimensional grid graph), created by random 2d walk

- a “layout”, created by assigning room types to nodes of abstract map

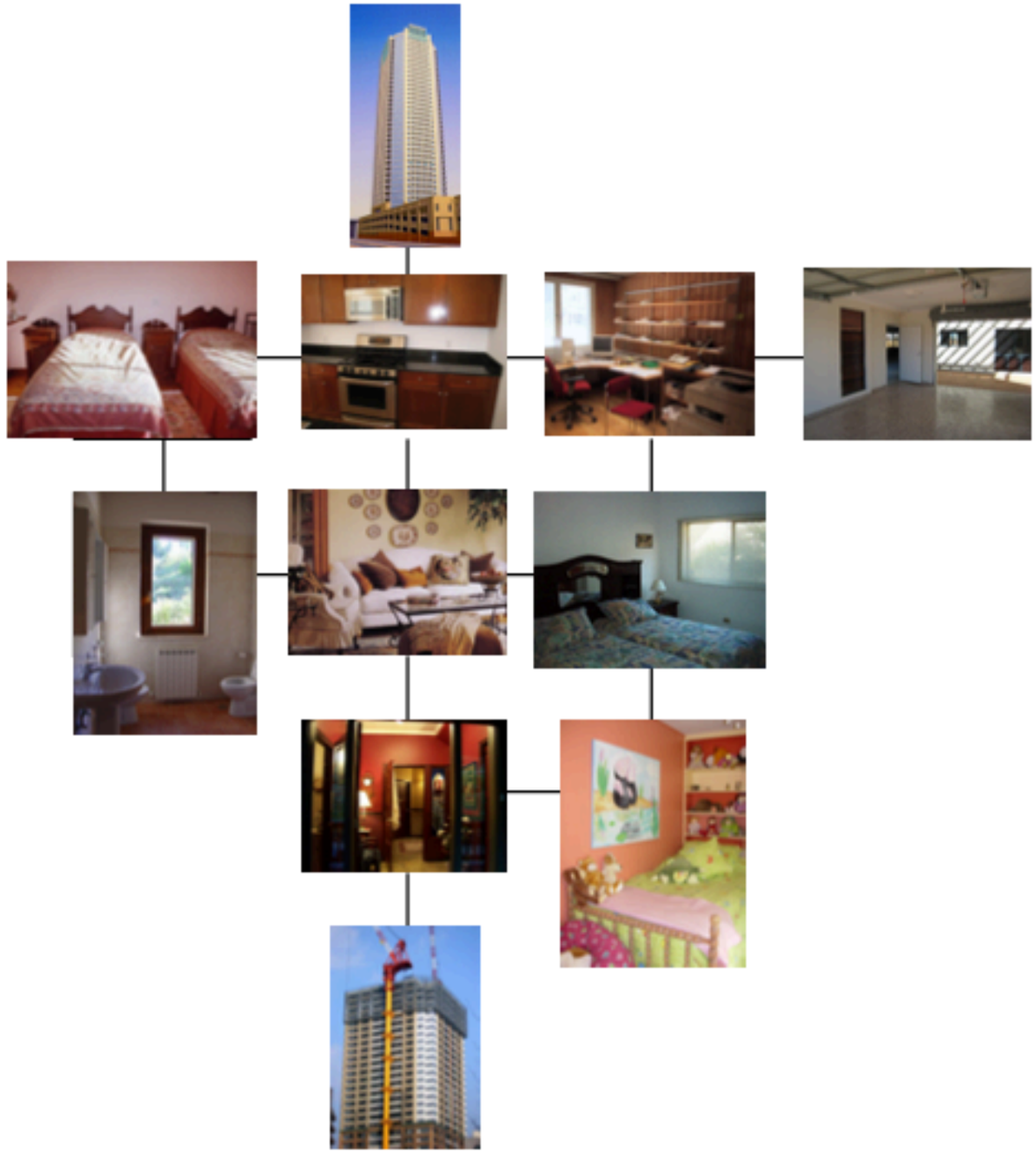


The MapWorld Environment

- an “abstract map” (connected subgraph of two-dimensional grid graph), created by random 2d walk

- a “layout”, created by assigning room types to nodes of abstract map

- a “game map”, created by sampling images of these types from ADE20k (Zhou *et al.* 2017)



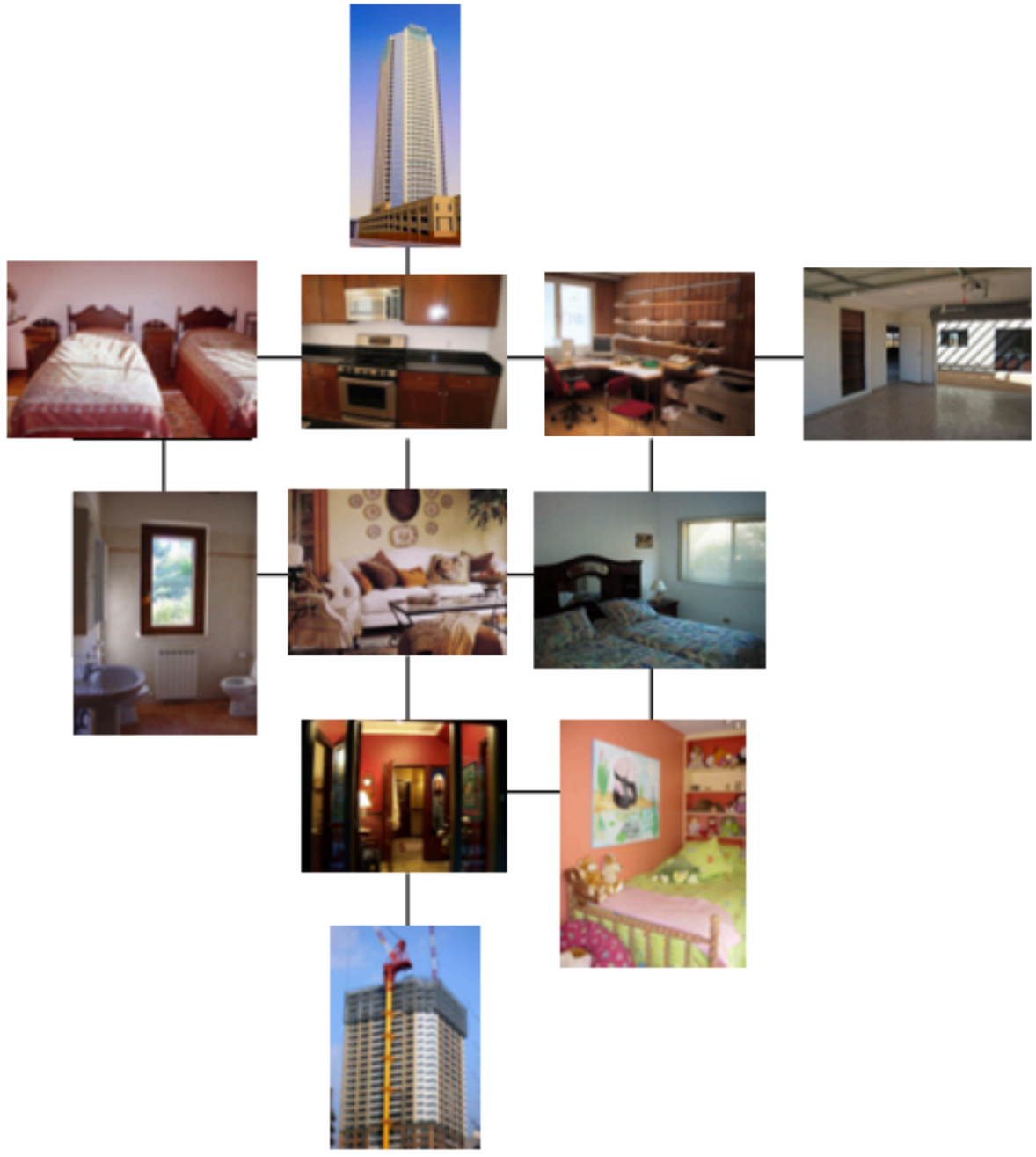
The MapWorld Environment

- an “abstract map” (connected subgraph of two-dimensional grid graph), created by random 2d walk

- a “layout”, created by assigning room types to nodes of abstract map

- a “game map”, created by sampling images of these types from ADE20k (Zhou *et al.* 2017)

- agents move from room to room, along cardinal dir.s



The MapWorld Environment

- an “abstract map” (connected subgraph of two-dimensional grid graph), created by random 2d walk

- a “layout”, created by assigning room types to nodes of abstract map

- a “game map”, created by sampling images of these types from ADE20k (Zhou *et al.* 2017)

- agents move from room to room, along cardinal dir.s

- related to recent glut of “house” environments (but much simpler)

[3]See for example (Savva et al., 2019; Adams et al., 2012; Johnson et al., 2016; Urbanek et al., 2019; Baroni et al., 2017a; Xia et al., 2018; Yan et al., 2018; Misra et al., 2018; Côté et al., 2018; Bennett and Shatkhin, 2018; Anderson et al., 2018; Savva et al., 2017; Gordon et al., 2017; Brodeur et al., 2017; Chang et al., 2017; Janarthanam and Lemon, 2011; Baroni et al., 2017b; Byron et al., 2007; Yamauchi et al., 2013).

The “Meetup in X” Game

- Two players. Their instructions: “Meet up in a room of type X.” (E.g., “meet up in a playroom.”)
  Success = Bonus.

- On MapWorld game board. 10 rooms, 4 of which are of target type. Players start at randomly selected location.

- Players can’t see each other, converse via chat. (Using *slurk*, (Schlangen *et al.* 2018) https://github.com/clp-research/slurk.)
  Signal when (they think they are) done.

- Game of *complete information* (rules & payoffs known), but *imperfect information* (only partial information about game state).
  Coordination game (payoffs coincide).

The `mux-2018` Dataset

- 430 usable dialogues collected on AMT. (Out of 547 started pairs…) Around $700 spent.

- 87% of those successful; in 10%, players in different rooms of correct type; 3% one player in wrong type.

- Average dialogue: 13.2 turns / 66.9 tokens / 165 secs / 28.3 navigation actions

Sync

Strategy

19	00:04	*GM (to B): You have to meet in: b/basement*
20	00:18	**A: i am in the bathroom**
21	00:29	**A: where are you at?**
22	00:33	**B: I'm outside. Let's move around a bit.**
23	00:38	**A: k**

Check

Strategy

49	01:01	*GM (to A): You can go: [/e]ast [/w]est*
50	01:11	**A: I am in the basement**
51	01:11	**B: I'm in a basement.**
52	01:23	**B: Mine has a white staircase**
53	01:28	**A: no**
54	01:37	**A: mine has wooden stair case**
55	01:55	**B: Okay. Should I try to move towards you?**
56	02:09	**A: Sure**
57	02:11	**B: Wooden? What else?**

Description

Check

66	02:27	**A: water heater and washer and dryer**
67	02:35	B (privately): s
68	02:35	*GM (to B): Nothing happened.You can go: [/n]orth [/e]ast*
69	02:40	B (privately): e
70	02:40	*GM (to B): url: /b/basement/ADE-train-00002494.jpg*
72	02:41	*GM (to B): You can go: [/e]ast [/w]est*
73	02:42	**A: a plastic chair and a screen door**
74	03:01	*GM: Attention: you are in the game for 3 minutes!*
75	03:18	**B: I'm there! I see the water heater, washer and dryer, sink, chair laying on top of the screen do**
76	03:26	B (privately): done
77	03:27	*GM: The '/done' command has been issued by you or your partner. To end the game, both players ne*
78	03:27	**A: yep**

B: i found one with black bottles

A: There is a light at the toop

B: maybe the one you were in

A: I left there

B: diamond shelves on top

A: I am in the one with the curved floor

B: ok, just find any of them

B: curved ceiling?

A: Curved wall and floor

A: A yellow light at the to[

A: *top

A: Go back to that room

B: are there racks and racks of bottles along tyhe hallway?

A: Go back to the room I described eariler

A: and I will go back there too

B: not sure whioch one that was

B: okay, found curved wall one

A

B

Medium sparrow, gray-brown upperparts, white underparts, black bib. Head has dark gray cap and sharply contrasting white eyebrow and cheek stripe. Bill is black. Long, round-tipped tail is edged with white. Legs and feet are gray. Forages on the ground and in low vegetation.

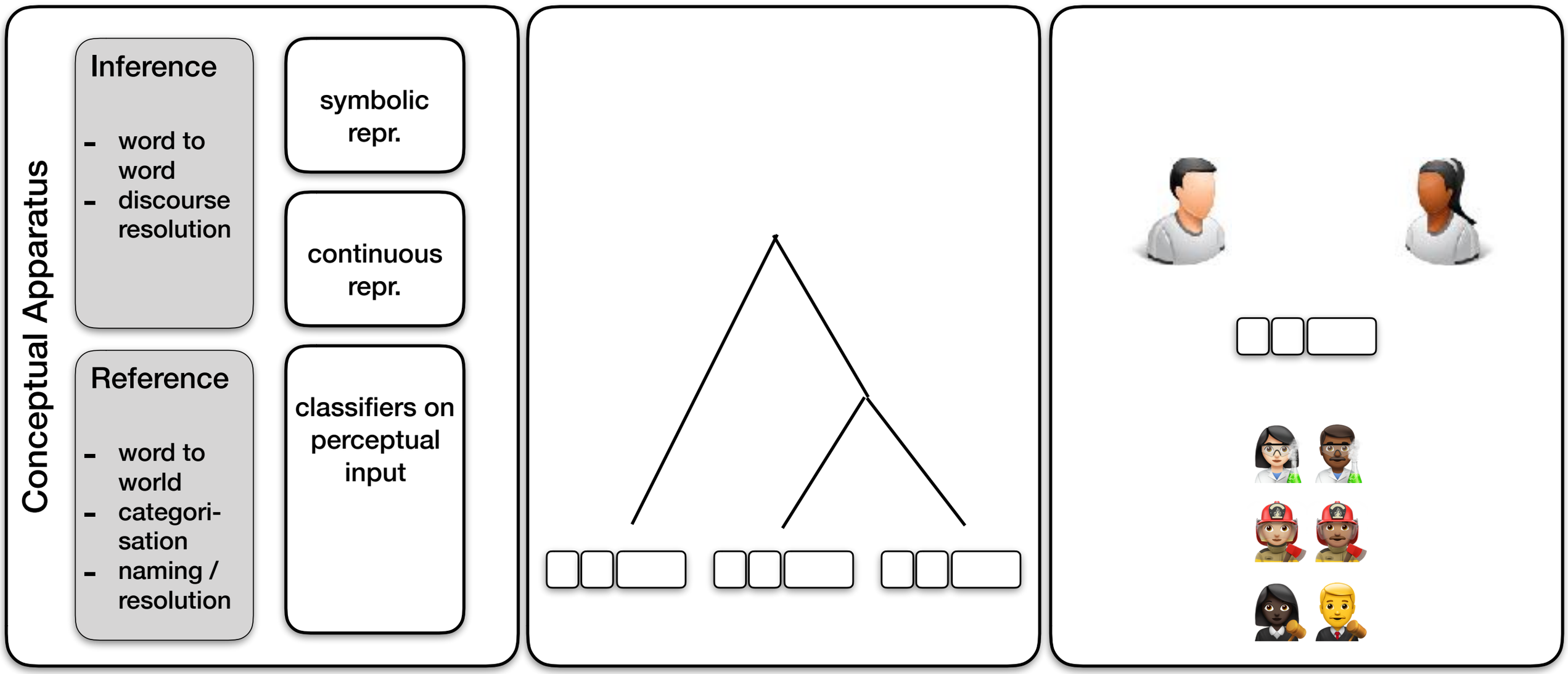
2019-09-24 01:07:16 - big - i think it is the one in the grass
2019-09-24 01:07:58 - big - what do you think?
2019-09-24 01:08:05 - unwavering - me too
2019-09-24 01:08:42 - unwavering - because of its white underparts
2019-09-24 01:09:05 - big - well, wait... the one says a black bib
2019-09-24 01:09:10 - big - so maybe the first one
2019-09-24 01:09:53 - big - thoughts?
2019-09-24 01:09:55 - unwavering - it does not have a long tail thou
2019-09-24 01:10:29 - big - i cant see the tail.....but the second doesnt have a bib.....but it does like vegetation
2019-09-24 01:10:49 - unwavering - so final answer should be ?
2019-09-24 01:11:07 - big - lets go withthe second one...in the grass
2019-09-24 01:11:12 - unwavering - ok
2019-09-24 01:11:44 - big - answer: The text describes a long tail, so we think it is the one in the grass
2019-09-24 01:11:44 - Cola Bot - The current proposal from big is **"The text describes a long tail, so we think it is the one in the grass"**
2019-09-24 01:11:44 - Cola Bot - Do you agree with your partner's answer? If not, please continue the discussion.
2019-09-24 01:12:14 - unwavering - agree

CoLa setting (Attari, Heckmann, Schlangen, sigdial 2019)

next steps, coordination part

- identify dialogue acts & strategies for conceptual coordination & update (see also Larsson, Noble 2019)

- goal: build agent that can justify its decisions and enter into discussions about them, deciding when to stand ground and when to yield



Make judgements using sub-symbolic knowledge and processes (classifiers on top of perceptual input; classifiers on top of continuous meaning representations).

Defend judgements using symbolic knowledge (explicit reasoning, explicit definitions).

Prediction: base categories cannot be defended, everything else must be?

functional *reprsnt.nal* *composition* *coordination*

Conceptual Apparatus

Inference

- word to word
- discourse resolution

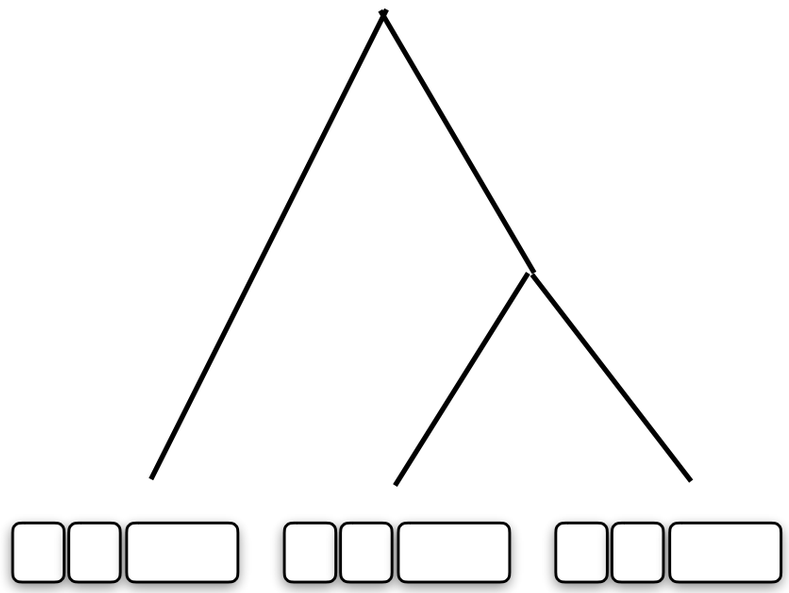
Reference

- word to world
- categori-sation
- naming / resolution

symbolic repr.

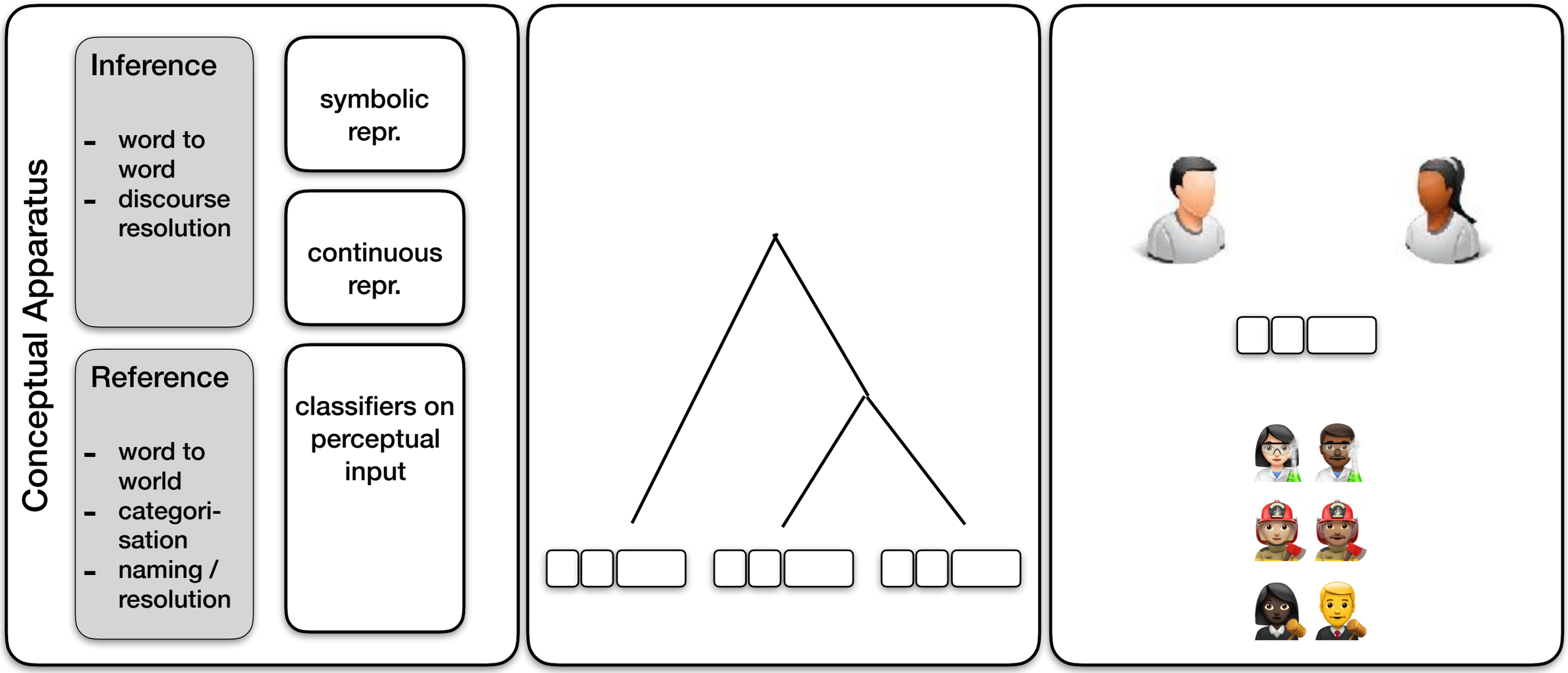
continuous repr.

classifiers on perceptual input



Combination of symbolic and subsymbolic processing by *parallelisation*… at the cost of multiplication of notions.. Have ⊨, ⊢ in classical versions (FOL) and in non-classical version (interpretation function is classifier, sets are fuzzy; entailment is judged by classifier on top of continuous representations)

Others: give up on classical interpretation of logical constants (Sadrzadeh); only have classical inference (Cooper, Larsson); project everything into one semantic space, no explicit reasoning (mainstream)



Many many many things left to do:

- get semantics with fuzzy sets / confidences to work, for judging statements, answering questions, …
- question answering, as area where inference and reference meet
- make syntax as shared basis more important to continuous representations
- properly integrating this into dialogue story: conceptual pacts, coordination

functional *reprsnt.nal* *composition* *coordination*

Conceptual Apparatus

Inference

- word to word
- discourse resolution

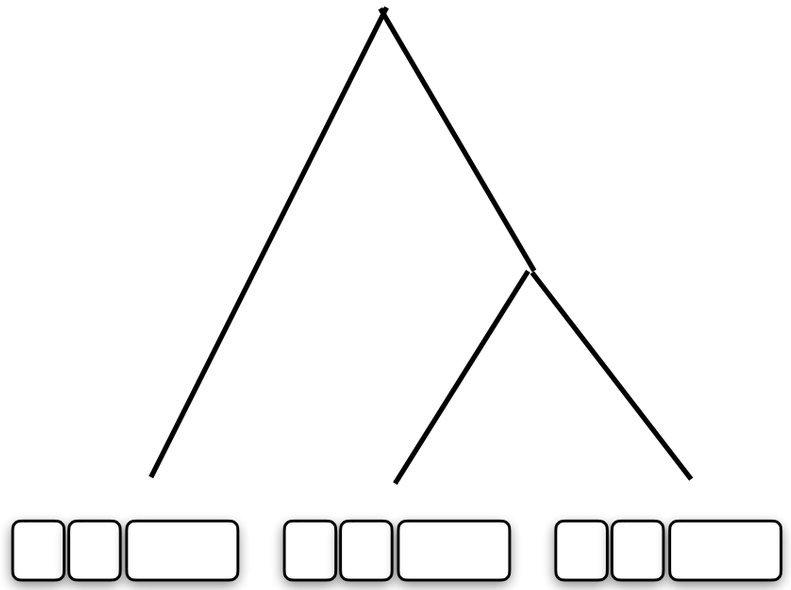
symbolic repr.

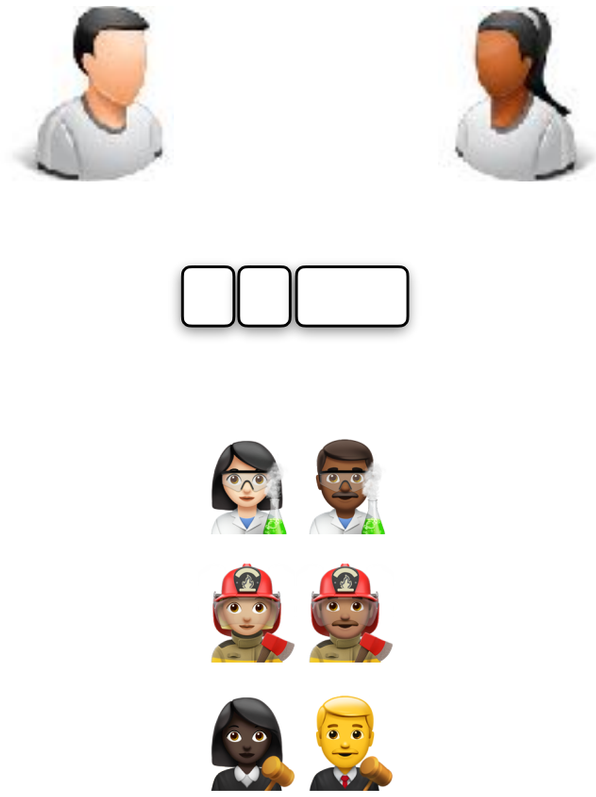
continuous repr.

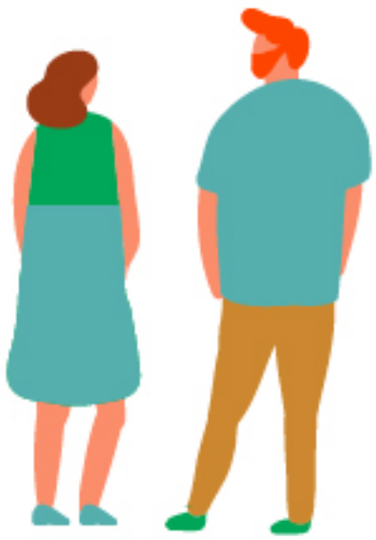
Reference

- word to world
- categori-sation
- naming / resolution

classifiers on perceptual input







Thank you.

Thanks also to my Bielefeld PhD students, Postdocs, and collaborators: Julian Hough, Sina Zarrieß, Casey Kennington, Nikolai Ilinykh, Soledad Lopez, Ting Han, Nazia Attari, Spyros Kousidis.

Funding received from CITEC, DFG.

References

References to our own work can be resolved via http://clp.ling.uni-potsdam.de/publications/ (where also the PDFs are available).
(First authors: Han, Kennington, Kousidis, Lopez, Schlangen, Zarrieß.)

- Andreas, J., Rohrbach, M., Darrell, T., & Klein, D. (2016). Neural Module Networks. In CVPR.
- Bowman, S. R., Potts, C., Angeli, G., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In EMNLP 2015.
- Bowman, S. R., Gauthier, J., Rastogi, A., Gupta, R., Manning, C. D., & Potts, C. (2016). A Fast Unified Model for Parsing and Sentence Understanding. In ACL 2016.
- Bruni, E., Boleda, G., & Baroni, M. (2012). Distributional Semantics in Technicolor. In ACL 2012 (pp. 136–145).
- Clark, E. V. (1987). The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), Mechanisms of Language Acquisition. Hilsdale, New Jersey, USA: Lawrence Erlbaum Associates.
- Elman, J. L. (1990). Finding Structure in Time. Cognitive Science, 14, 179–211.
- Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930-1955. In Studies in linguistic analysis.
- Goldberg, Y., & Levy, O. (2014). word2vec Explained: Deriving Mikolov et al.'s Negative-Sampling Word-Embedding Method. arXiv preprint arXiv:1402.3722.
- Goodman, N. (1955). Fact, Fiction, & Forecast, Cambridge Massachusetts: Harvard University Press
- Haber, J., Baumgärtner, T., Takmaz, E., Gelderloos, L., Bruni, E., & Fernández, R. (2019). The PhotoBook Dataset: Building Common Ground through Visually-Grounded Dialogue. In ACL 2019.
- Harris, Z. S. (1954). Distributional Structure. Word, 10(2–3), 146–162.

References

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1–32.
- Kaji, N., & Kobayashi, H. (2017). Incremental Skip-gram Model with Negative Sampling. In EMNLP (pp. 363–371).
- Kruszewski, G., & Baroni, M. (2015). So similar and yet incompatible : Toward automated identification of semantically compatible words. In The 2015 Annual Conference of the North American Chapter of the ACL (pp. 964–969).
- Lampert, C. H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, 951–958.
- McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. Behavior Research Methods, Instruments & Computers, 37(4), 547--559.
- Silberer, C., & Lapata, M. (2014). Learning Grounded Meaning Representations with Autoencoders. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 721–732.
- Socher, R., Lin, C. C., Ng, A. Y., & Manning, C. D. (2011). Parsing Natural Scenes and Natural Language. In Proc. 28th International Conference on Machine Learning.
- Udagawa, T., & Aizawa, A. (2019). A Natural Language Corpus of Common Grounding under Continuous and Partially-Observable Context. In AAAI.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention Is All You Need. In NIPS.