A Joint Model of Language and Perception for Grounded Attribute Learning

Cynthia Matuszek, Nicholas FitzGerald,
Liefeng Bo, Luke Zettlemoyer, Dieter Fox

SE 367

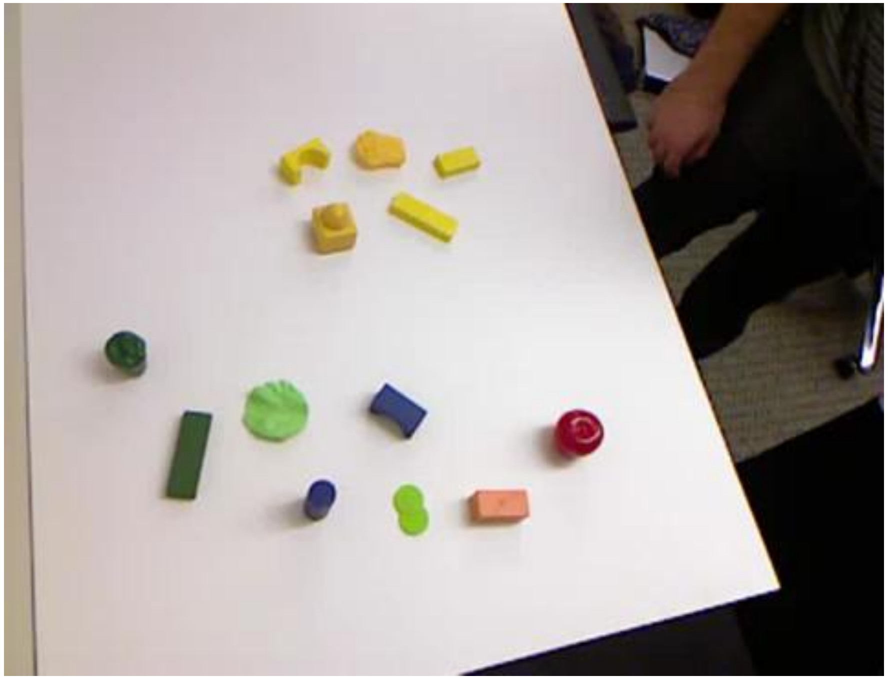
Debidatta Dwibedi

The Vision

- Robots should learn about their environment by interacting with humans
  - Not by being programmed by them!
- Problems:
  - Tough for the layman to ‘teach’ a robot
  - Inability of the robot to make inductions
- Solutions:
  - Point to object and describe in natural language
  - Use language and perception to ground attributes like colors and shapes

Objective

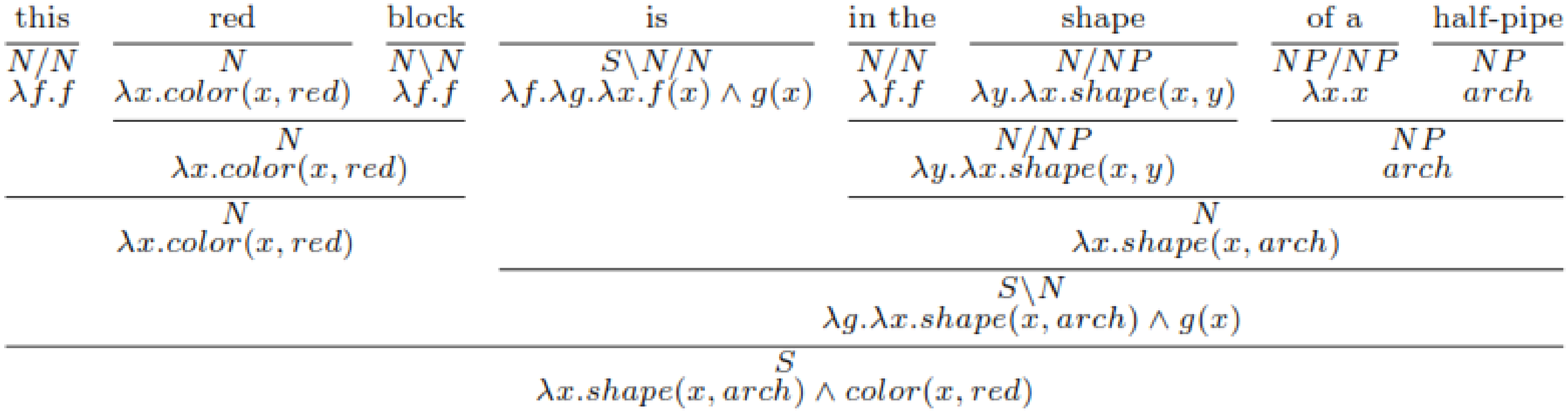
- Select objects based on attribute
- Learn **previously unknown** attributes
  - Yellow: new word describing new idea



“Which are the yellow objects?”

Semantic Parsing

- To produce the robot’s (mental?) representation
- Combinatory Categorial Grammars *[Steedman (book) 2000, Kwiatkowski et al 2010, 2011]* used to **parse sentences** into **lambda calculus expressions**



this — N/N — $\lambda f.f$

red — N — $\lambda x.color(x, red)$

block — $N\backslash N$ — $\lambda f.f$

is — $S\backslash N/N$ — $\lambda f.\lambda g.\lambda x.f(x) \wedge g(x)$

in the — N/N — $\lambda f.f$

shape — N/NP — $\lambda y.\lambda x.shape(x, y)$

of a — NP/NP — $\lambda x.x$

half-pipe — NP — $arch$

N — $\lambda x.color(x, red)$

N/NP — $\lambda y.\lambda x.shape(x, y)$

NP — $arch$

N — $\lambda x.color(x, red)$

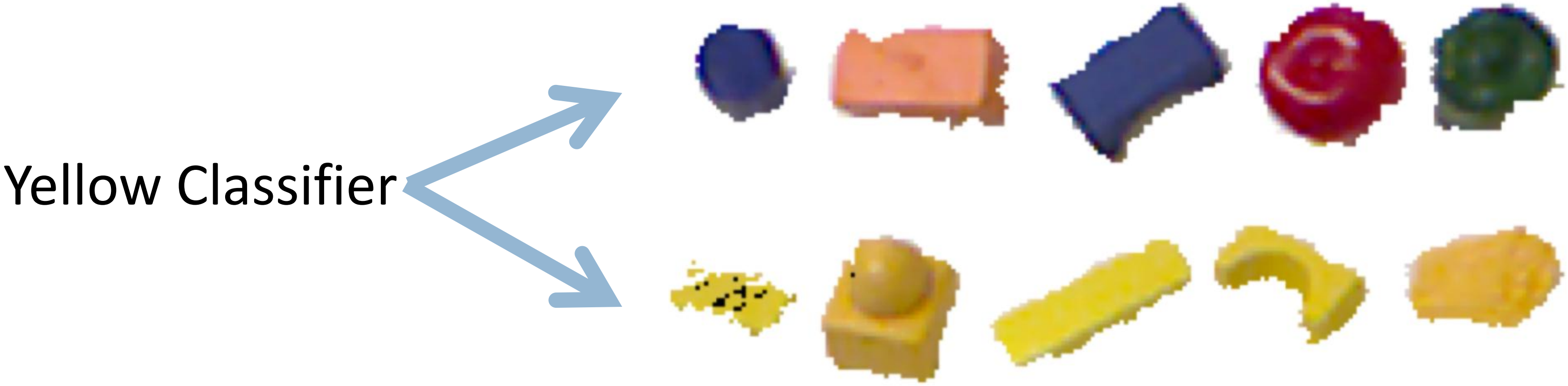
N — $\lambda x.shape(x, arch)$

$S\backslash N$ — $\lambda g.\lambda x.shape(x, arch) \wedge g(x)$

S — $\lambda x.shape(x, arch) \wedge color(x, red)$

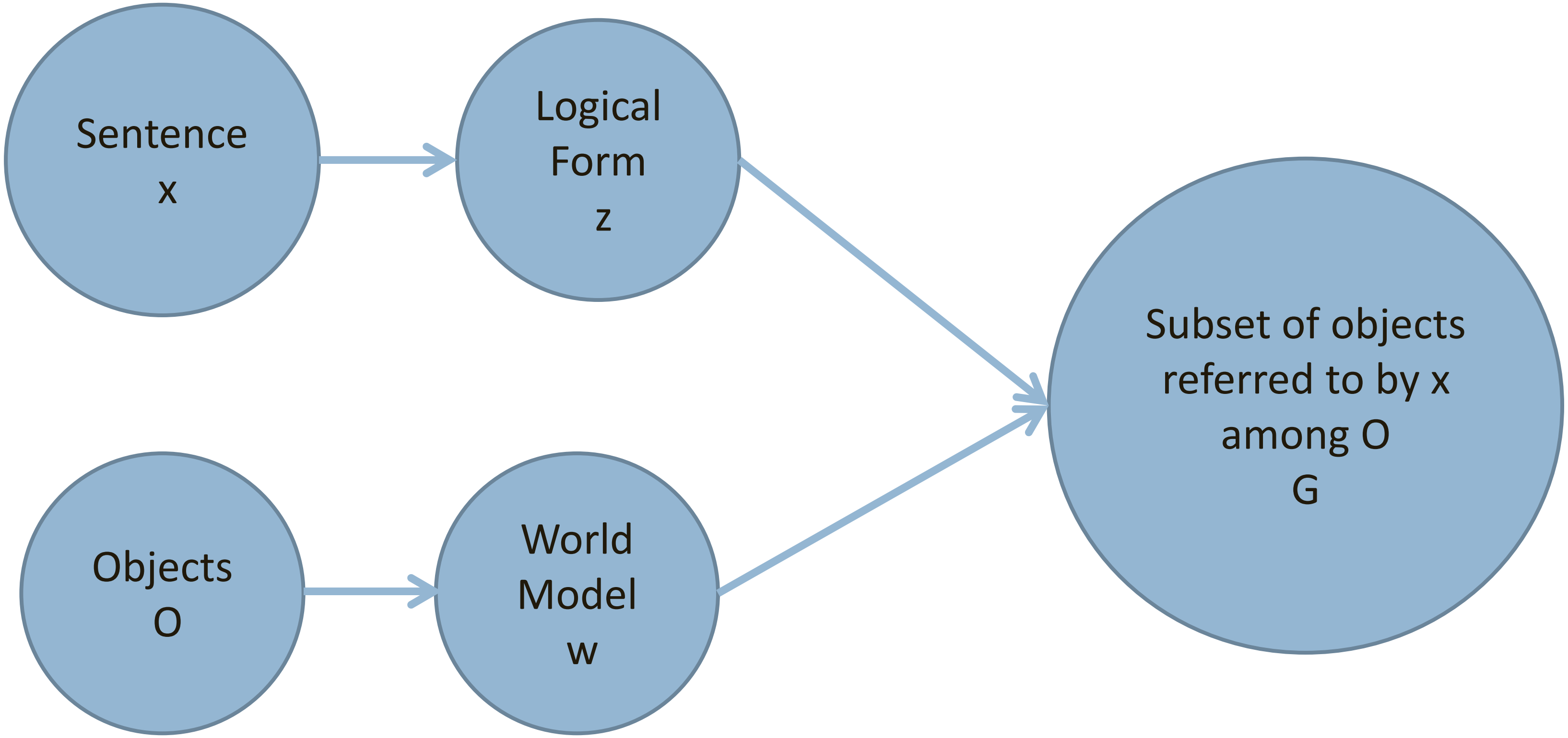
Perceptual Model

- Segment objects from environment
- Set of binary classifiers
  - each perceptual classifier is applied independently
  - use logistic regression to train classifiers on colour and shape features

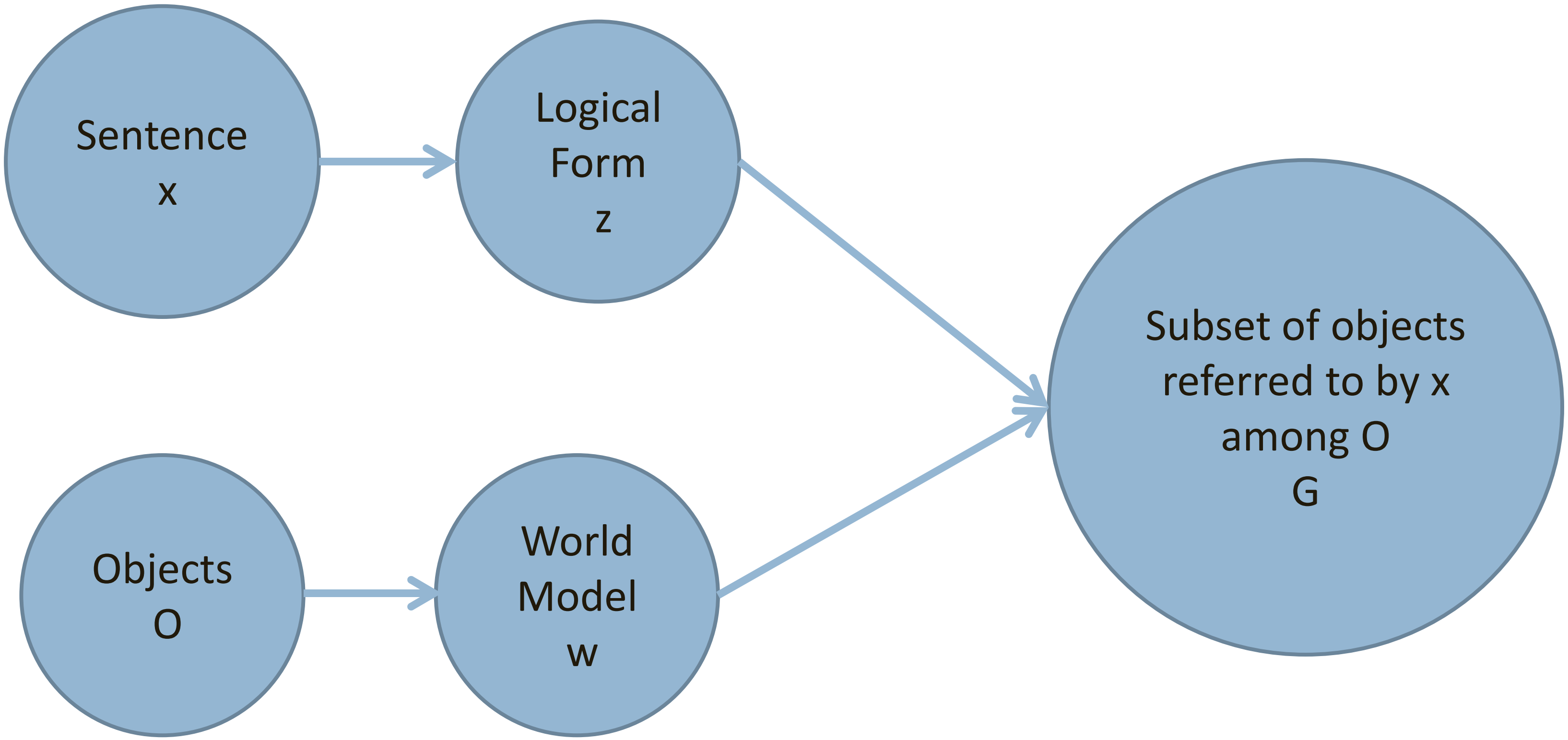


Yellow Classifier

Joint Model


Sentence
x
Logical
Form
z
Objects
O
World
Model
w
Subset of objects
referred to by x
among O
G

Joint Model



Sentence x

Logical Form z

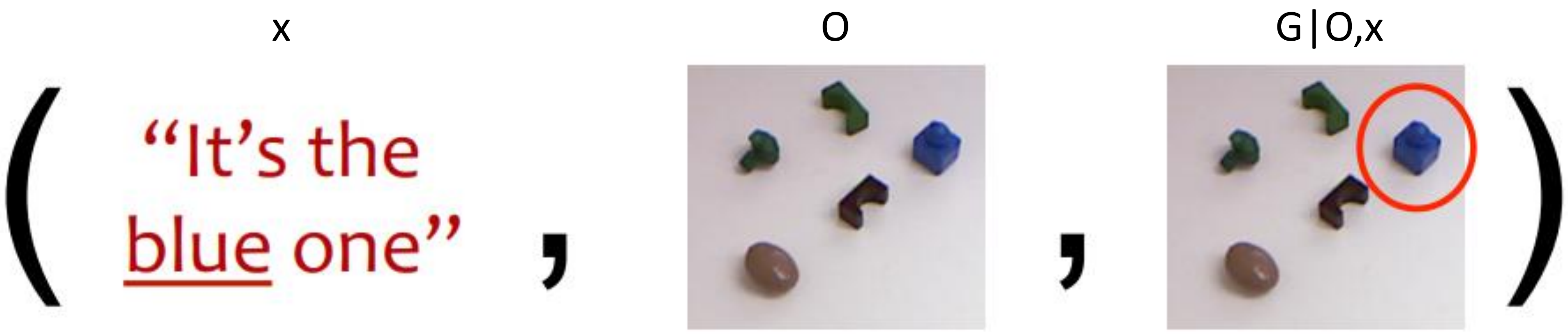
Objects O

World Model w

Subset of objects referred to by x among O G

$$\underbrace{P(G, z, w \mid x, O)}_{\text{Joint Probability}} = \underbrace{P(z \mid x)}_{\text{Parsing Model}}\underbrace{P(w \mid O)}_{\text{World Model}}\underbrace{P(G \mid z, w)}_{\text{Grounding Query}}$$

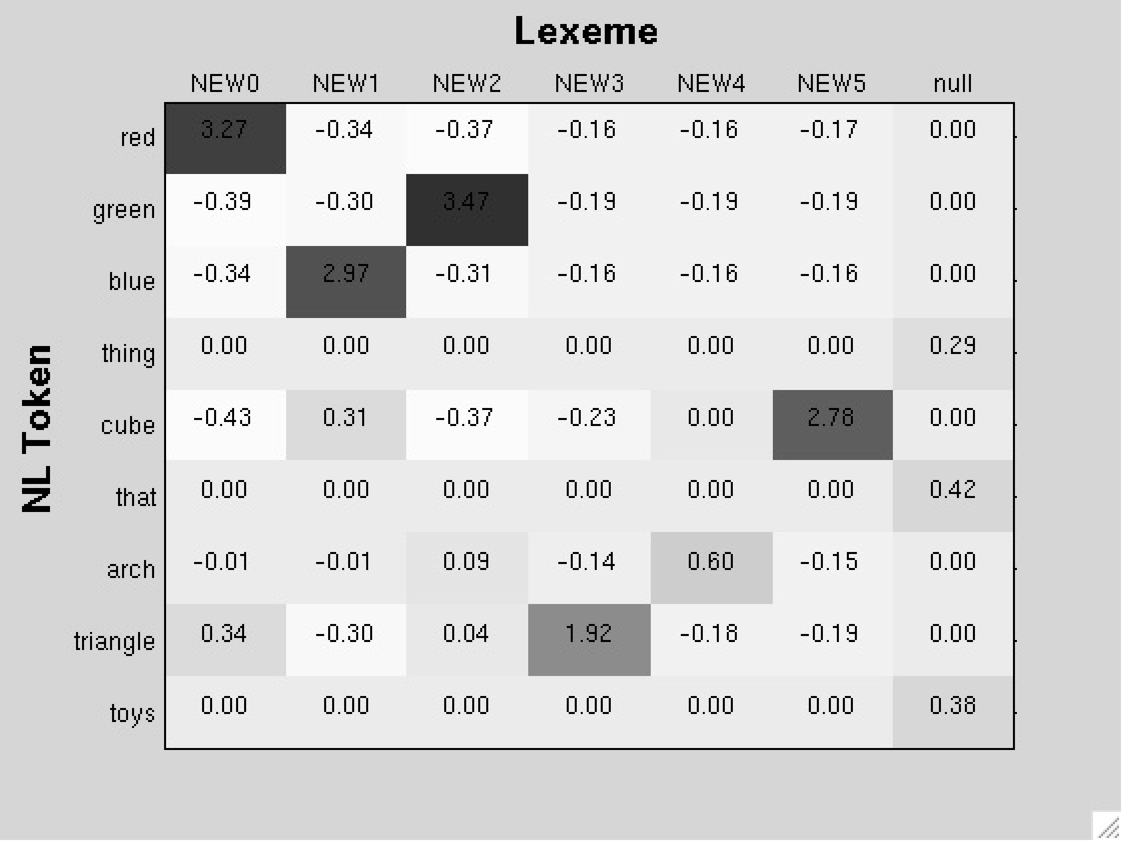
Unsupervised Learning



(“It’s the blue one” , ,)

- Initialization
  - Train an initial supervised model from labeled scenes
- Learn new attributes
  - Found N new attributes
  - Add *N* new, unknown attribute classifiers
  - Initialize to a small, near-uniform distribution
  - Pair with every unknown word/phrase
  - Expectation Maximization

Results

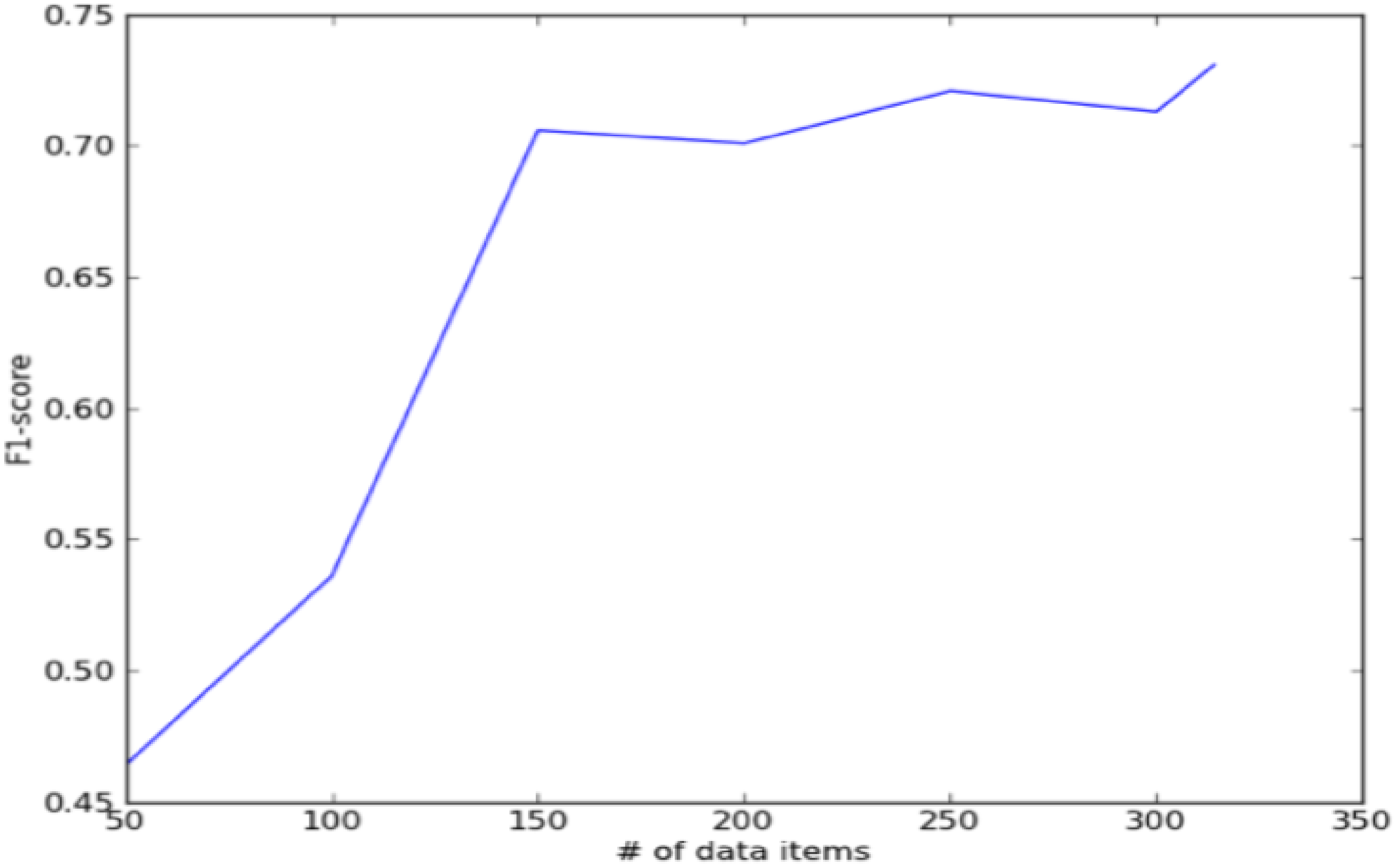


Lexeme

NL Token

	NEW0	NEW1	NEW2	NEW3	NEW4	NEW5	null
red	3.27	-0.34	-0.37	-0.16	-0.16	-0.17	0.00
green	-0.39	-0.30	3.47	-0.19	-0.19	-0.19	0.00
blue	-0.34	2.97	-0.31	-0.16	-0.16	-0.16	0.00
thing	0.00	0.00	0.00	0.00	0.00	0.00	0.29
cube	-0.43	0.31	-0.37	-0.23	0.00	2.78	0.00
that	0.00	0.00	0.00	0.00	0.00	0.00	0.42
arch	-0.01	-0.01	0.09	-0.14	0.60	-0.15	0.00
triangle	0.34	-0.30	0.04	1.92	-0.18	-0.19	0.00
toys	0.00	0.00	0.00	0.00	0.00	0.00	0.38

Results


0.75
0.70
0.65
0.60
0.55
0.50
0.45
F1-score
50
100
150
200
250
300
350
# of data items

Not all Humans are good Teachers

- Since people were told to describe the objects being pointed to in the manner they would do it to an infant, some descriptions are not helpful in learning attributes:
  - *“This object is a fake piece of green lettuce. Do not try to eat!”* (Unexpected input)
  - *“This is a toy”* (no attributes mentioned)
  - *“This is a rectangular block”* when the block was cylindrical (Wrong descriptions due to noisy data or otherwise)

References

- *A Joint Model of Language and Perception for Grounded Attribute Learning*(2012) Cynthia Matuszek and FitzGerald, N. and Zettlemoyer, L. and Bo, L. and Fox, D.