Pragmatically Informative Image Captioning with Character-Level Inference

Reuben Cohn-Gordon
Christopher Potts, Noah Goodman
NAACL 2018

What caption describes B1?

a vehicle

B1

What caption describes B1 **and not B2**?

~~*a vehicle*~~

B1

B2

What caption describes B1 **and not B2**?

bus

B1

B2

What caption describes B1 **and not B2**?

~~bus~~

B1

B2

What caption describes B1 **and not B2**?

red bus

B1

B2

What caption describes B1 **and not B2**?

~~red bus~~

B1

B2

What caption describes B1 **and not B2**?

British bus

B1

B2

What caption describes B1 **and not B2**?

~~British bus~~

B1

B2

What caption describes B1 **and not B2**?

old British bus

B1

B2

Overview

B1

B2

Overview

B1

B2

Bayesian models of referential captioning

Overview

B1

B2

Bayesian models of referential captioning

Applied to neural image captioning

Overview

B1

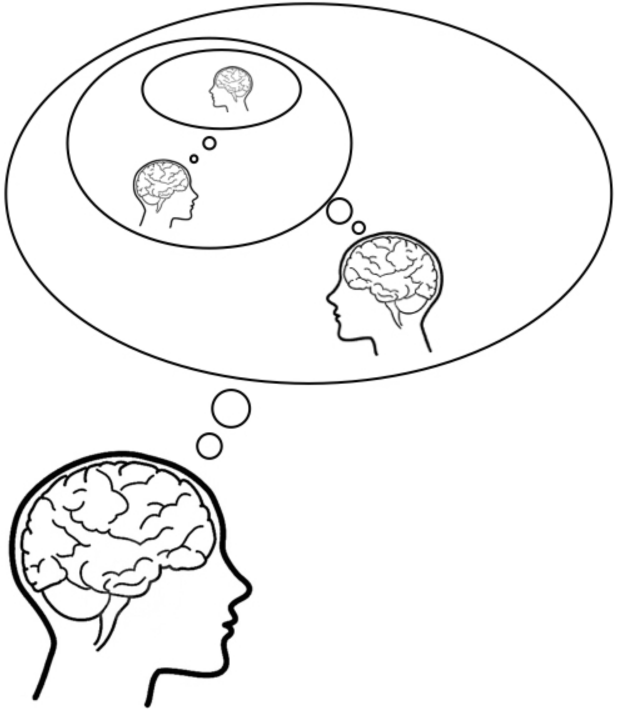
B2

Bayesian models of referential captioning

Applied to neural image captioning

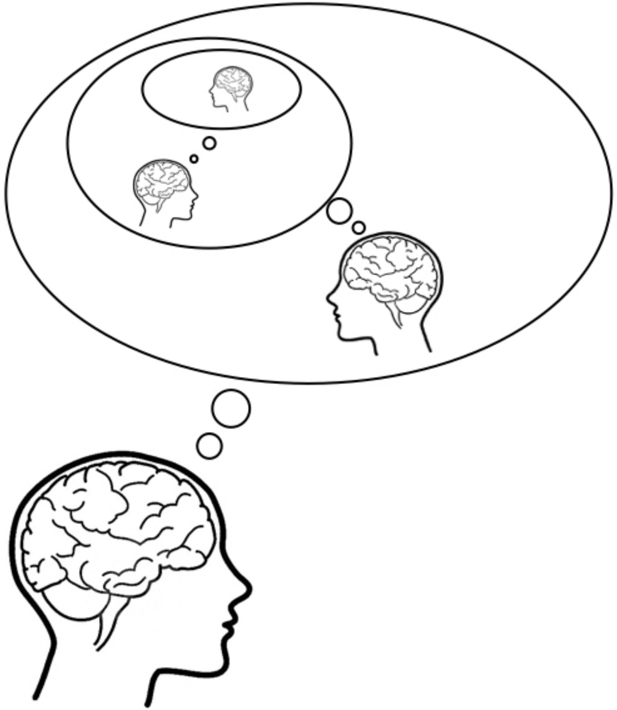
With character-level inference (my contribution)

Bayesian Models of Reference



http://www.glascherlab.org/social-decisionmaking

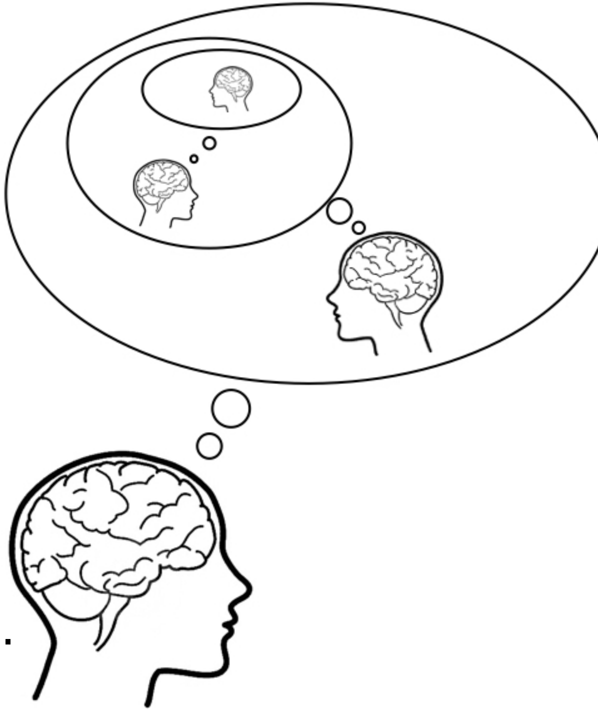
Bayesian Models of Reference



http://www.glascherlab.org/social-decisionmaking

Rational Speech Acts (RSA) paradigm as formalism

Bayesian Models of Reference

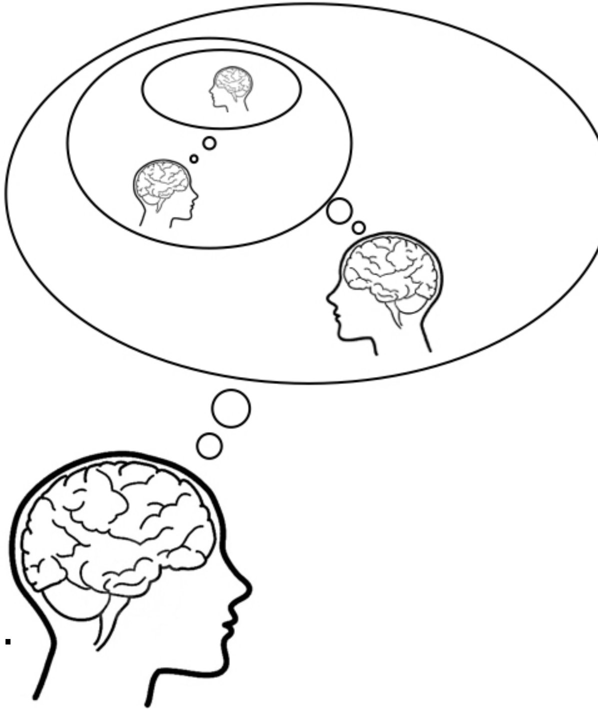


http://www.glascherlab.org/social-decisionmaking

Rational Speech Acts (RSA) paradigm as formalism

Speaker reasons about listener reasoning about speaker...

Bayesian Models of Reference



Rational Speech Acts (RSA) paradigm as formalism

Speaker reasons about listener reasoning about speaker...

Speakers and listeners are **conditional distributions**

http://www.glascherlab.org/social-decisionmaking

S_0	bus	red bus
B1	1/2	1/2
B2	1	0

B1

B2

S_0	bus	red bus
B1	**1/2**	**1/2**
B2	**1**	**0**

B1

B2

L_0	**B1**	**B2**
bus	**1/3**	**2/3**
red bus	**1**	**0**

$$L_0(image|caption) = \frac{S_0(caption|image) \cdot P(image)}{\Sigma_{\textbf{i'in image}} S_0(caption|i') \cdot P(i')}$$

S_0	bus	red bus
B1	**1/2**	**1/2**
B2	**1**	**0**

B1

B2

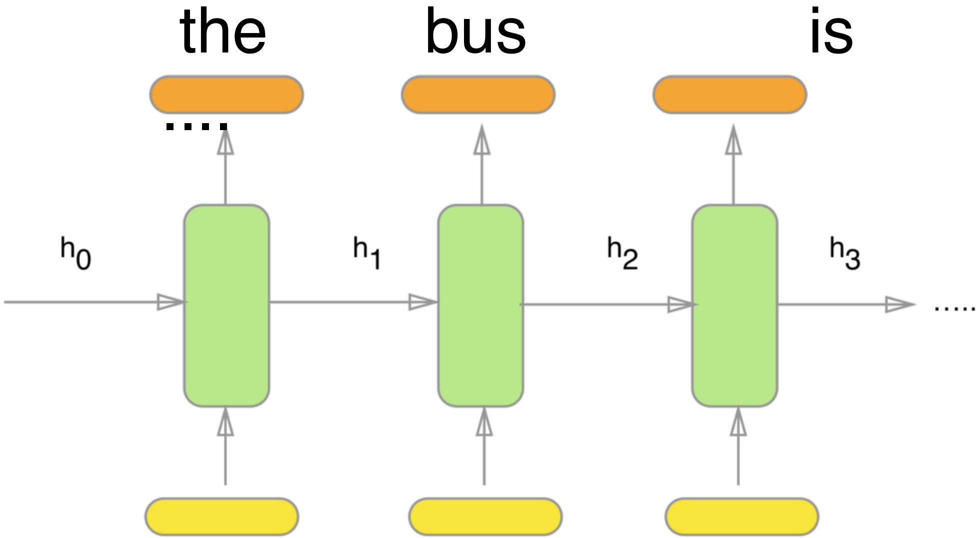
L_0	**B1**	**B2**
bus	**1/3**	**2/3**
red bus	**1**	**0**

$$L_0(image \mid caption) = \frac{S_0(caption \mid image) \cdot P(image)}{\sum_{i' \text{ in image}} S_0(caption \mid i') \cdot P(i')}$$

S_1	bus	red bus
B1	**1/4**	**3/4**
B2	**1**	**0**

$$S_1(caption \mid image) = \frac{L_0(image \mid caption)^a \cdot P(caption)}{\sum_{c' \text{ in captions}} L_0(image \mid c')^a \cdot P(c')}$$

Neural Image Captioning



a	0.7
the	0.21
...	...

a	bus	0.5
	truck	0.2
	road	0.1
	fish	0.001
	...	...

a	bus	<Stop>	0.9
		and	0.01
		bicycle	...
		banana	
		...	

a	bus	<Stop>	0.9
			0.01
			...

RSA for Natural Language Processing

S_0: a **neural captioner** *p(caption|image)*

Then the S_1 **does not require training** on a dataset of referential captions

Utterance set: **all combinations of words** (up to some length)

See (Mao et al., 2016a, Vedantam et al., 2017) for applications to captioning
See: Dale and Reiter, 1995, Monroe and Potts, 2015, Andreas and Klein, 2016, Monroe et al., 2017

The Issue of the Utterance Space

$$S_1(\textit{caption}|\textit{image}) = L_0(\textit{image}|\textit{caption})^a \cdot P(\textit{caption}) \,/\, \left(\sum_{\mathbf{c' in\ captions}} L_0(\textit{image}|c')^a \cdot P(c') \right)$$

The Issue of the Utterance Space

$$S_1(\textit{caption}|\textit{image}) = L_0(\textit{image}|\textit{caption})^a \cdot P(\textit{caption}) \,/\, \left(\sum_{\mathbf{c' in\ captions}} L_0(\textit{image}|c')^a \cdot P(c') \right)$$

Intractable sum: $\sum_{\mathbf{c'\ in\ captions}}$

What if we do pragmatics at each timestep?

TARGET

DISTRACTOR

S_0:	S_1:

TARGET

DISTRACTOR

	S_0:	S_1:
red	0.3	0.5
bus	0.6	0.1
...	...	...

TARGET

	S_0:	S_1:
red bus	0.7	0.8
is	0.2	0.1
	...	...

DISTRACTOR

TARGET

	S_0:	S_1:
red bus …		
is …		

DISTRACTOR

Pragmatics at the Level of Characters

Pragmatics at the Level of Characters

Pragmatics at the Level of Characters

- Let’s apply the incremental approach to a character RNN!

Pragmatics at the Level of Characters

- Let’s apply the incremental approach to a character RNN!

Pragmatics at the Level of Characters

- Let's apply the incremental approach to a character RNN!

- Global number: 30^{60} = 42391158275216203514294433201000000000000000000000000000000000000000000000000000000000000

Pragmatics at the Level of Characters

- Let's apply the incremental approach to a character RNN!

- Global number: 30^{60} = 42391158275216203514294433201000000000000000000000000000000000000000000000000000000000000

Pragmatics at the Level of Characters

- Let’s apply the incremental approach to a character RNN!

- Global number: 30^{60} = 42391158275216203514294433201000000000000000000000000000000000000000000000000000000000000

- Local number: 30x60 = 1800

Pragmatics at the Level of Characters

- Let’s apply the incremental approach to a character RNN!

- Global number: 30^{60} = 42391158275216203514294433201000000000000000000000000000000000000000000000000000000000000

- Local number: 30x60 = 1800

Pragmatics at the Level of Characters

- Let’s apply the incremental approach to a character RNN!

- Global number: 30^{60} = 42391158275216203514294433201000000000000000000000000000000000000000000000000000000000000

- Local number: 30x60 = 1800

- (Advantageous even considering only captions with non-zero probability)

TARGET

DISTRACTOR

r

TARGET

DISTRACTOR

re

TARGET

DISTRACTOR

red

TARGET

DISTRACTOR

		S_0	S_1
red	**d**	0.27	0.90
	b	0.50	0.08

TARGET

DISTRACTOR

red do

TARGET

DISTRACTOR

red dou

TARGET

DISTRACTOR

red doub

TARGET

DISTRACTOR

red doubl

TARGET

DISTRACTOR

red double

TARGET

DISTRACTOR

red double d

TARGET

DISTRACTOR

red double de

TARGET

DISTRACTOR

red double dec

TARGET

DISTRACTOR

red double deck

TARGET

DISTRACTOR

red double decke

TARGET

DISTRACTOR

red double decker

TARGET

DISTRACTOR

red double decker b

TARGET

DISTRACTOR

red double decker bu

TARGET

DISTRACTOR

red double decker bus

TARGET

DISTRACTOR

r
t
a
d
o
b
l

g
c
p
i
m
h
s
e
j
w
n
f
u
q
x
^
z
k
y
&

In practice, we use a beam search

DISTRACTOR

TARGET

re
th
a
do
tw
bu
or
ar
op
lu
tr
ta
an
au
r
la
le
de
gr
ve
a
ot
oc
ol
pa
ap
ir
lo

In practice, we use a beam search

DISTRACTOR

TARGET

red
the
a r
a d
dou
two
bus
ora
ope
arr
lus
thi
tre
tal
re
lar
an
a b
gre
dee
aut
ver
a p
are
oth
a
old
ano

In practice, we use a beam search

DISTRACTOR

TARGET

red
the
a re
a do
doub
two
bus
ther
oran
reda
open
arri
this
lush
tall
tree
red
larg
an o
gree
a bu
auto
deer
ared
othe
tre
a r
anot

In practice, we use a beam search

DISTRACTOR

TARGET

red d
red b
a red
the r
the b
red f
a dou
red p
doubl
red s
red c
the d
red m
red a
orang
two d
two r
therr
red t
two g
this
red g
red u
tall
arris
bus i
two s
redal

In practice, we use a beam search

DISTRACTOR

TARGET

red do
red bu
a red
the re
red br
the bu
a doub
red dr
double
red pu
red fo
red su
red fu
red co
the do
orange
red fl
the br
two re
two do
red fr
red an
therre
red fi
two gr
red pi
red tr
red ma

In practice, we use a beam search

DISTRACTOR

TARGET

red dou
red bus
a red d
the red
the bus
red bri
a doubl
red bro
double
red dre
red col
red dru
red pub
a red b
red sub
red fur
the dou
red pur
red fol
red bui
orange
red foo
red flo
the bri
red dri
two red
red and
red fou

In practice, we use a beam search

DISTRACTOR

TARGET

red doub
red bus
a red do
the red
the bus
red bric
red doup
red doug
a double
red broc
double d
red dour
red colo
red doum
red dres
red publ
red drum
the doub
red purp
red fold
red sub
a red bu
red buil
red food
the bric
two red
red fure
red flow

In practice, we use a beam search

DISTRACTOR

TARGET

red doubl
red bus i
red bus o
a red dou
the bus i
red brick
the red d
red doupl
a double
red dough
red bus w
the red b
red broch
double de
red bus d
red doube
red bus s
red bus p
red color
red dress
red bus r
red drum
red bus a
the doubl
red dourr
red publi
red bus v
red bus g

In practice, we use a beam search

DISTRACTOR

TARGET

red double
red doubled
red bus on
a red doubl
red bus is
red doubler
red doubly
red bus in
the bus is
the red dou
red brick b
a double de
red bus wit
red douple
double deck
the red bus
red doughnu
red doubley
red doubles
red bus par
red bus sto
red colored
red bricked
red douber
red bus doo
red douber
red bus doo
red brick r

In practice, we use a beam search

DISTRACTOR

TARGET

red double d
red doublede
a red double
red doublerd
red doubly d
red bus on t
the red doub
red doubled
red doubledd
red bus is o
red brick bu
a double dec
red doubledr
the bus is r
red bus with
red bus on a
red bus in t
red bus on s
red douple d
red bus is p
double decke
red bus on r
red doughnut
red bus in a
the red bus
red doubley
red bus in r
red bus is i

In practice, we use a beam search

DISTRACTOR

TARGET

red double de
red doubledec
a red double
red bus on th
red doublerd
red doubleder
the red doubl
red doubly do
red bus is on
red doubled d
a double deck
red brick bui
the bus is re
red doubly da
red bus with
red bus on a
red doubledra
red doubledd
red bus in th
red doubleded
red bus is pa
red bus on ro
a red doubled
red douple do
red bus on st
double decker
red doubly de
red doughnut

In practice, we use a beam search

DISTRACTOR

TARGET

red double dec
a red double d
red doubledeck
red bus on the
red doubledere
the red double
red doubly dou
red bus is on
red doubled do
a double decke
red brick buil
red doublerd d
the bus is red
red bus in the
red doubly dar
red doubleded
red bus on roa
red bus is par
red bus on str
double decker
red bus on a s
red doublerd b
a red doublede
red douple doo
the red bus is
red doubledeca
red doublerd o
red doubledeco

In practice, we use a beam search

DISTRACTOR

TARGET

red double deck
a red double de
red bus on the
red doubledeck
red doublederes
the red double
red doubly doub
red brick build
a double decker
red bus is on t
red doubled doo
red doublerd do
red bus in the
red doubly dark
red bus on road
red bus is park
red bus on stre
red doublerd bu
red bus on a st
a red doubledec
red douple door
red doubledecke
the red bus is
double decker b
red doubleded b
red doubledeco
red doublerd re
red doubledecat

In practice, we use a beam search

DISTRACTOR

TARGET

red double decke
red double deck
a red double dec
red bus on the s
the red double d
red doublederes
red doubly doubl
red doubledeck b
red brick buildi
a double decker
red bus is on th
red doubled door
red doubledeck d
red doublerd doo
red bus is parke
red bus on stree
red doublerd bus
red bus on the r
red bus on a str
red doubledecked
red doubly darke
double decker bu
red doubly dark
red doubledeck t
red doubleded bu
red bus in the r
red doubly dark
red doubledeck t

In practice, we use a beam search

DISTRACTOR

TARGET

red double decker
red double decked
a red double deck
red bus on the st
the red double de
red double deck b
red doubly double
red doublederes o
red double deck p
red doubledeck bu
red brick buildin
red bus is on the
red doubledeck do
red double deck r
a double decker b
red doublerd door
red bus is parked
red bus on street
red bus on the ro
red bus on a stre
red doubled doors
red doubled doorw
red doubledecked
red doubly darker
red double deck d
double decker bus
red doublerd bus
red double decket

In practice, we use a beam search

DISTRACTOR

TARGET

red double decker
red double decked
a red double decke
red bus on the str
the red double dec
red double deck bu
red doubly double
red brick building
red bus is on the
red doublederes on
red doubledeck bus
red double deck re
red doubledeck doo
a double decker bu
red double deck pa
red bus on the roa
red bus on a stree
red double deckere
red doublerd doors
red double deck pu
red double deck do
red doubled doors
red doubled doorwa
red double decket
red bus is parked
red doubledecked b
double decker bus

In practice, we use a beam search

DISTRACTOR

TARGET

red double decker b
red double decked b
red double decker r
red double decker d
red double decker p
red double decked r
red double decked p
a red double decker
red double decker c
red bus on the stre
the red double deck
red double decked c
red double deck bus
red double decked d
red doubly double d
red double decker o
red double decker a
red doublederes on
red double decker s
red double deck red
red doubledeck door
red double decker g
red bus is on the s
a double decker bus
red double decker w
red brick building
red double decked t
red bus on the road

In practice, we use a beam search

DISTRACTOR

TARGET

red double decker bu
red double decked bu
red double decker re
red double decker do
red double decked re
red double decker br
a red double decker
red double decked pu
red double decker pu
red bus on the stree
the red double decke
red double decked do
red double decker pa
red doubly double de
red double decker co
red double decked ci
red double deck red
red double decker ci
red double decker gr
red double decker or
red double decker wi
red bus is on the st
red double decked pi
red bus is on the si
red double decker st
red double decker be
red double decker ca

In practice, we use a beam search

DISTRACTOR

TARGET

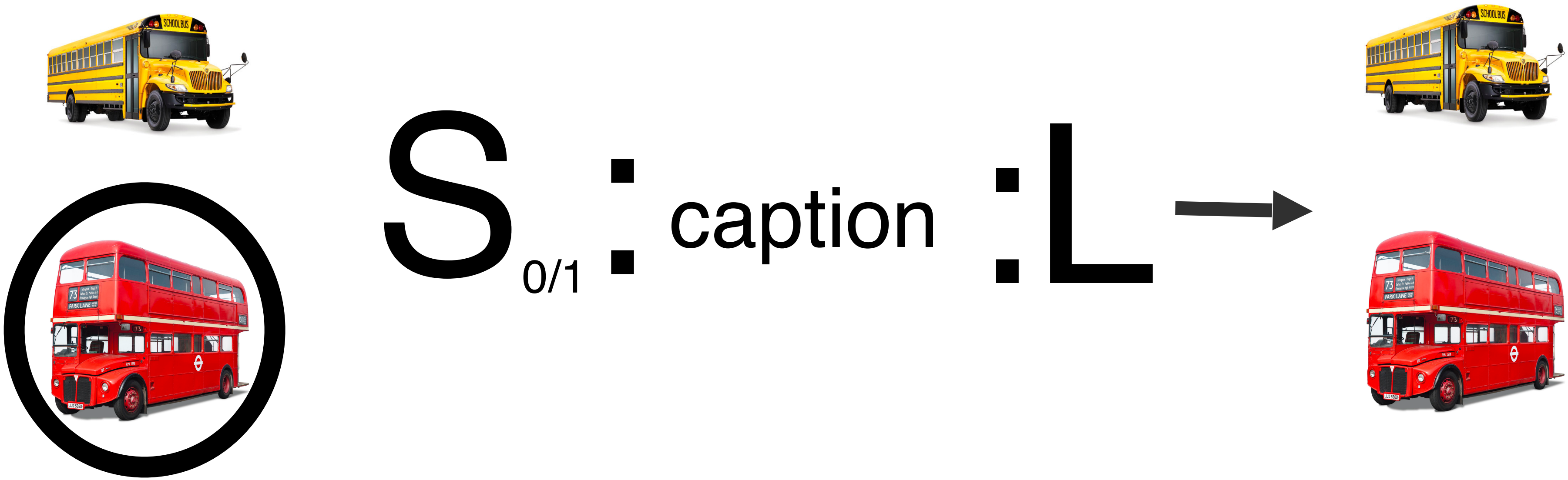
red double decker bus
red double decked bus
red double decker red
red double decker dou
red double decked red
red double decker bro
a red double decker b
red bus on the street
the red double decker
red double decked dou
red double decked pub
red doubly double dec
red double decker pub
red double decker par
red double decker doo
red bus is on the st
red double decked pi
red bus is on the si
red double decker st
red double decker be
red double decker ca
red double decker bus
red double decked bus
red double decker red
red double decker dou
red double decked red
red double decker bro
red double decker sto

In practice, we use a beam search

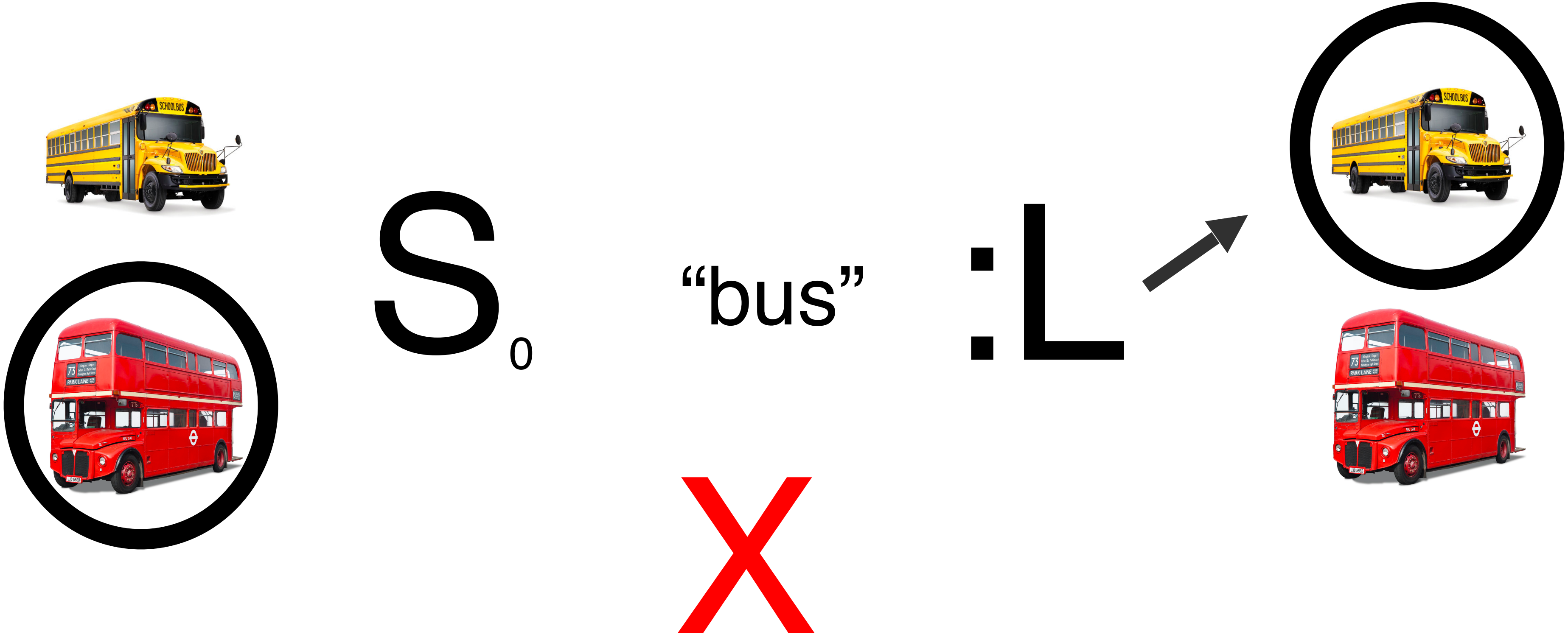
DISTRACTOR

TARGET

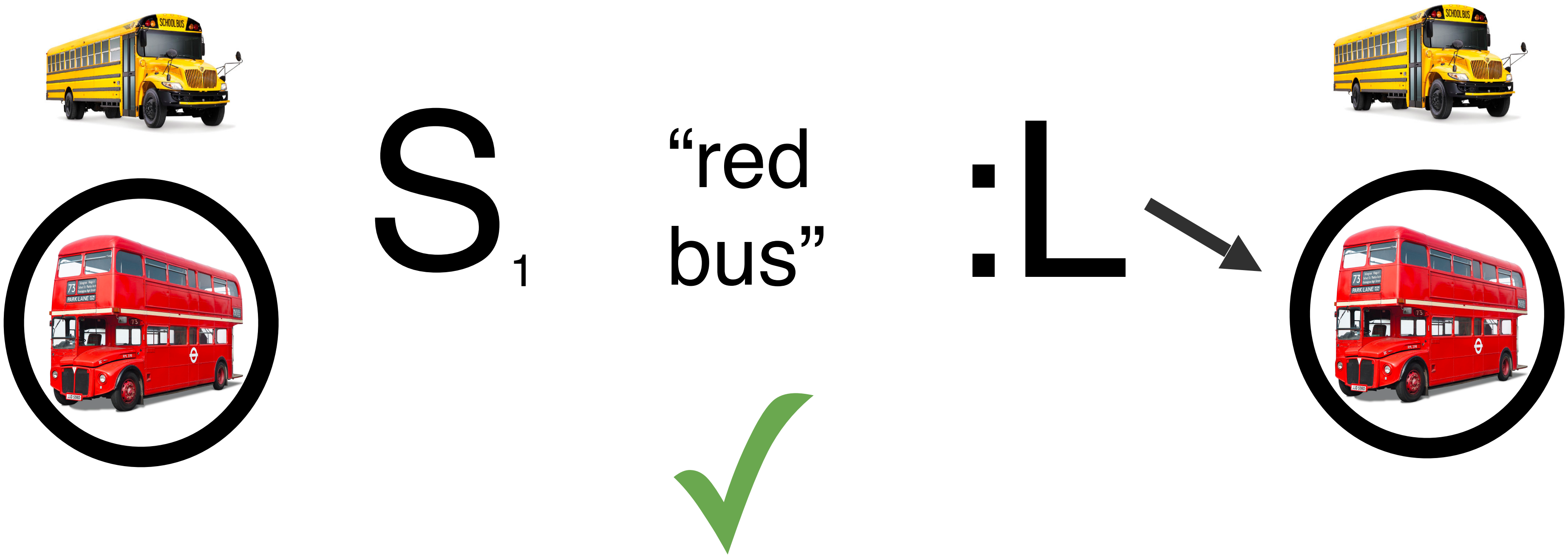
Evaluation



Evaluation



Evaluation



Evaluation

- Is **separately trained listener** more likely to pick target given S_0 or S_1 captions?

- This method **does not require a dataset of pragmatically informative captions:** just clusters of similar images

Results

Model	TS1 Accuracy	TS2 Accuracy

Results

- **S_1 is better than S_0**

Model	TS1 Accuracy	TS2 Accuracy
Char S_0	48.9	47.5
Char S_1	<u>68.0</u>	<u>65.9</u>

Results

- **S_1 is better than S_0**
- **Character-level is better than word-level!**

Model	TS1 Accuracy	TS2 Accuracy
Char S_0	48.9	47.5
Char S_1	<u>68.0</u>	<u>65.9</u>
Word S_0	57.6	53.4
Word S_1	60.6	57.6

The Bigger Picture

- Incremental pragmatics can yield global pragmatic effects
- Realistic utterance spaces are not a barrier to Bayesian pragmatics
- Bayesian pragmatics for natural language beyond referential image captioning

Incremental S_1

$S_0(\textit{word} \mid \textit{partial\_cap}, \textit{img})$

$$L_0(\textit{img} \mid \textit{partial\_cap}, \textit{word}) = \frac{S_0(\textit{word} \mid \textit{partial\_cap}, \textit{img})P(\textit{img})}{\sum_{i'} S_0(\textit{word} \mid \text{partial\_cap}, i')P(i')}$$

$$S_1(\textit{word} \mid \textit{partial\_cap}, \textit{img}) = \frac{L_0(\textit{img} \mid \textit{partial\_cap}, \textit{word})^a S_0(\textit{word} \mid \textit{partial\_cap}, \textit{img})}{\sum_{w'} L_0(\textit{image} \mid \textit{partial\_caption}, w')^a S_0(w' \mid \textit{partial\_cap}, \textit{img})}$$

Referential

B1

B2

A good caption for an image is **not just true**, but also *pragmatically informative*

A *pragmatically informative* caption allows a listener to **identify the target image**.

Which image does this caption refer to?

red bus

B1

B2

Which image does this caption refer to?

red bus

B1

B2

Which image does this caption refer to?

bus

B1

B2

Which image does this caption refer to?

bus

B1

?

B2