THE UNIVERSITY OF CHICAGO

NLP representations from a perspective of human cognition

Allyson Ettinger

CLSP, Johns Hopkins University

Nov 18 2019

The big goal

- NLP is trying to solve “natural language understanding”
- This can be defined in various ways
- Ideal: achieve human capacity to extract, represent, and deploy information from language input

How to assess “understanding”?

- How do we assess the *information* that a system has captured?
- Downstream tasks?

How to assess “understanding”?

- Current challenge in NLP: powerful pre-trained models are beating our current benchmarks
- But no one really thinks we have mastered “understanding”
- This is a mismatch that needs to be addressed
- Our dominant question: how can we better understand and more effectively evaluate what our models actually “know” about language

Using human cognition as a lens

- We’re going to examine this from the perspective of human cognition
- *What do we need humans for? Planes don’t flap their wings …*
- Concept of understanding is defined based on humans
- Essentially all NLP benchmarks use human judgments at some level

What about humans to aspire to

- Certain levels of human understanding make sense for us to emulate with our systems – specifically, endpoint of comprehension

- Other aspects (errors, early stages) not clear we want to emulate

- But sometimes our models do resemble these other aspects
- Worth identifying, thinking about why this is happening, and determining what needs to change to target the endpoint of comprehension

Outline

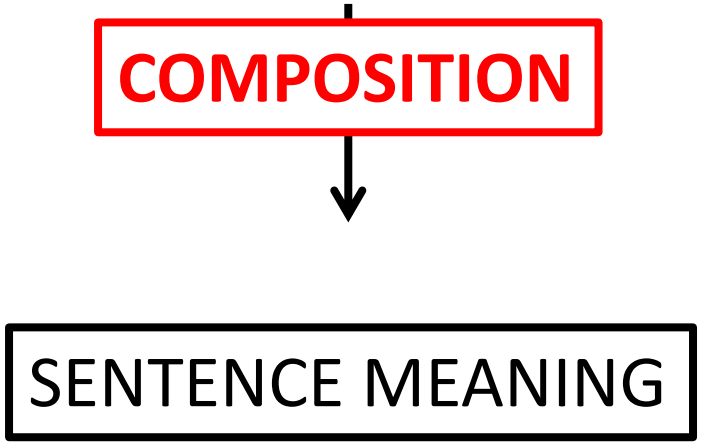
1. Assessing systematic composition in sentence encoders
2. Simpler models as approximation of real-time predictive response
3. Evaluating pre-trained LMs against human predictive responses

Outline

1. **Assessing systematic composition in sentence encoders**
2. Simpler models as approximation of real-time predictive response
3. Evaluating pre-trained LMs against human predictive responses

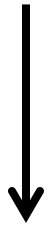
Learning sentence representations

The turquoise giraffe recited the sonnet but did not forgive the flight attendant



How are we doing at meaning composition?

The turquoise giraffe recited the sonnet but did not forgive the flight attendant



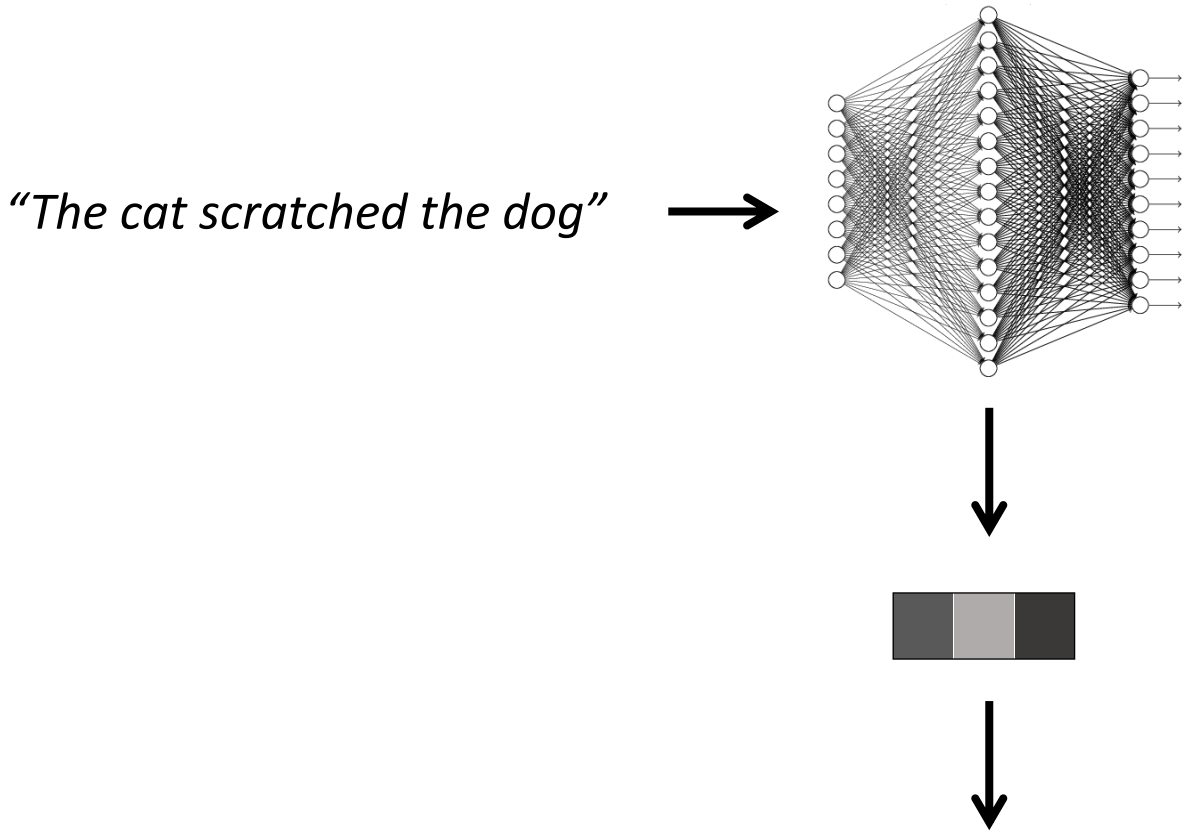
??

[.23 -.04 .45 .13 …]

Probing tasks

Is my sentence encoder capturing word content?

“The cat scratched the dog”



Is “dog” in the sentence?

Probing tasks

- Ettinger et al. (2016), Adi et al. (2016)

- Dates back over a decade in neuroscience: multivariate pattern analysis, Haxby et al. (2001)

Our work

- Target aspects of sentence meaning relevant to composition
- Additional measures to control tests and increase confidence in conclusions

Control 1:

Control 1: sentence generation

"*professor* = AGENT of *help*"

The professor helped the student

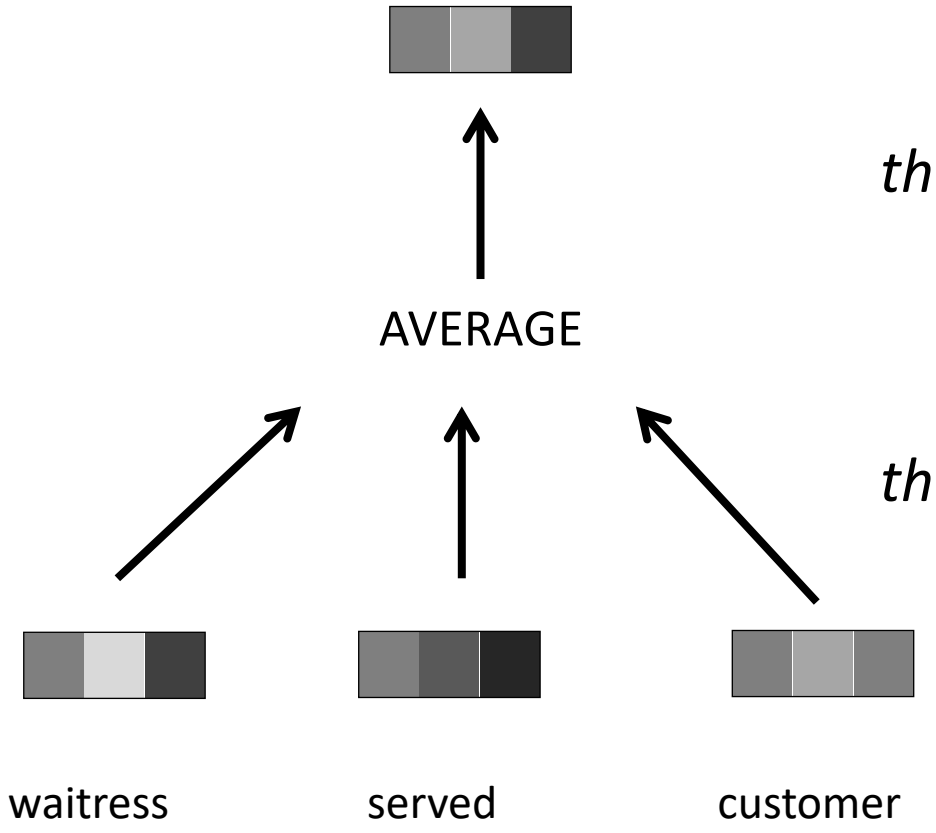
The professor is not helping the executive

The lawyer is being helped by the professor

The professor that the girl likes helped the man

Control 2:

Control 2: Bag-of-words check

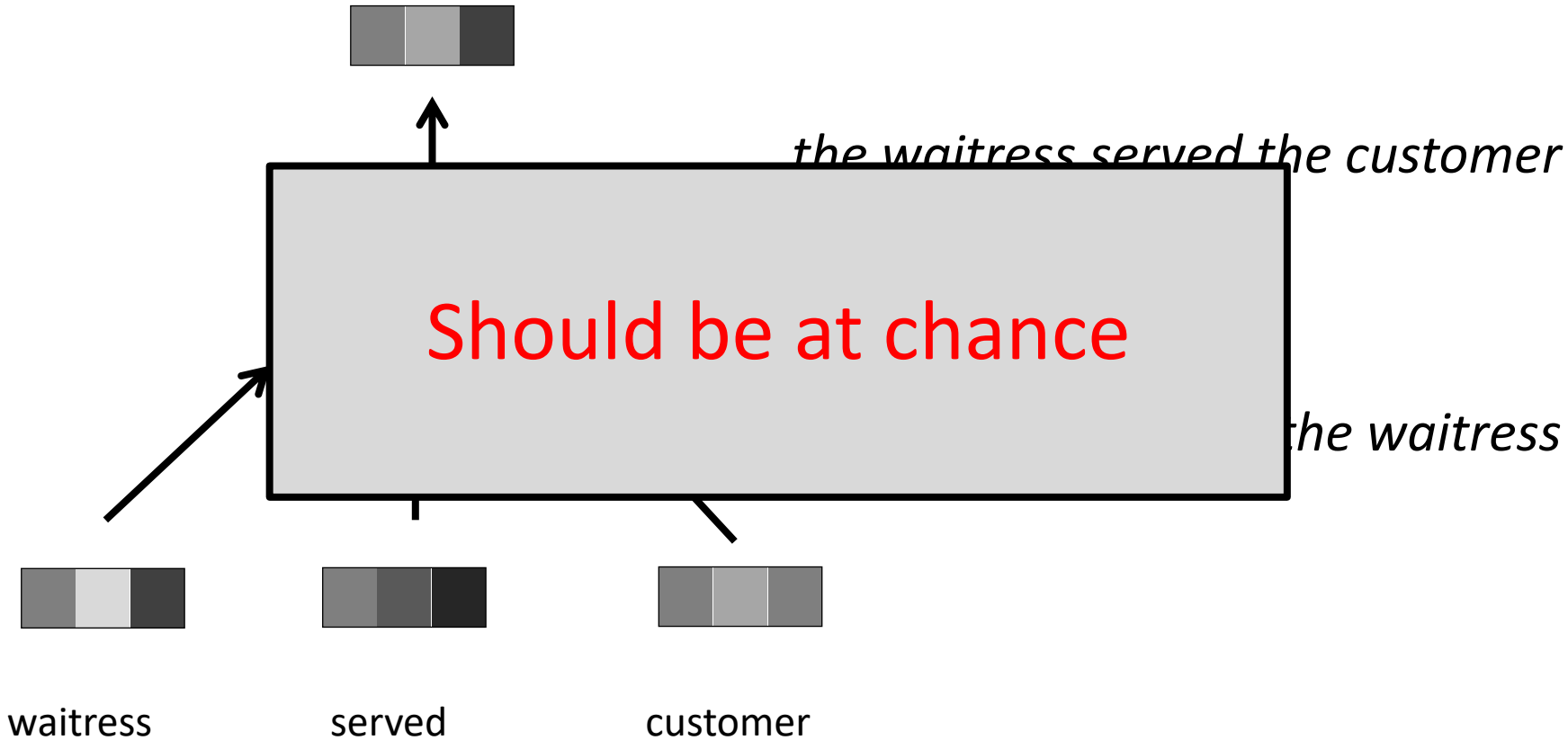


the waitress served the customer

??

the customer served the waitress

Control 2: Bag-of-words check



- What information do we know that humans extract systematically?

Target information types

- **Semantic role** (who did what to whom?)

- **Negation** (what happened and what didn’t?)

Semantic role: is *x* agent of *y*?

SENT: *The **waitress** who **served** the customer is sleeping*

X-PROBE: *waitress*

Y-PROBE: *serve*

LABEL: +1

SENT: *The waitress who served the **customer** is **sleeping***

X-PROBE: *customer*

Y-PROBE: *sleep*

LABEL: -1

Negation: did *y* happen?

SENT: *The waitress is serving the customer who is **not** actually **sleeping***

Y-PROBE: *sleep*

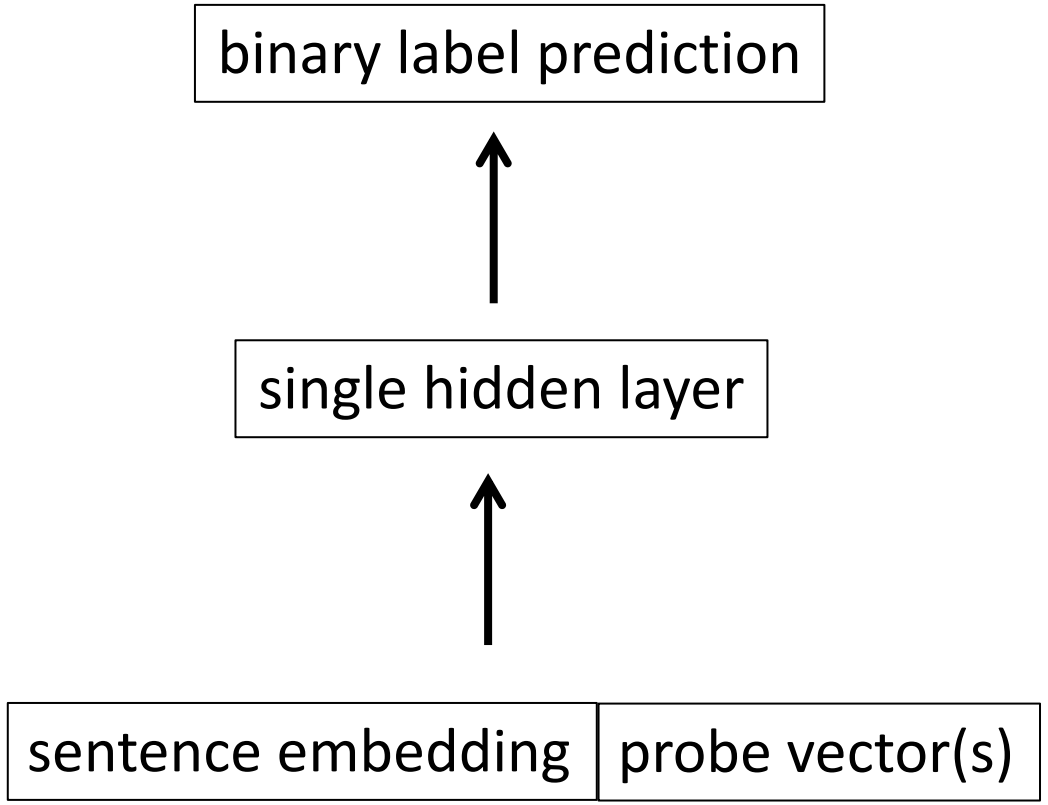
LABEL: -1

SENT: *The waitress is **not** actually serving the customer who is **sleeping***

Y-PROBE: *sleep*

LABEL: +1

MLP classifier



Sentence embedding models

- BOW: Bag-of-words vector averaging
- SDAE: Sequential Denoising Autoencoder (Hill et al., 2015)
- ST-UNI, ST-BI: SkipThought – uniskip and biskip (Kiros et al., 2015)
- InferSent (Conneau et al. 2017)

- 2400 dimensions

Sanity check: surface tasks

- **word content**
  - given probe *x*: is *x* present in sentence?
- **word order**
  - given probes *x*, *y*: does *x* precede *y* in sentence?

Adi et al., 2016

Classification accuracy

CONTENT	ORDER	ROLE	NEG

Classification accuracy

	CONTENT	ORDER	ROLE	NEG
BOW	100.0	55.0	51.3	50.9

Classification accuracy

	CONTENT	ORDER	ROLE	NEG
BOW	100.0	55.0	51.3	50.9
SDAE	100.0	92.9		
ST-UNI	100.0	93.2		
ST-BI	96.6	88.7		
InferSent	100.0	86.4		

Classification accuracy

	CONTENT	ORDER	ROLE	NEG
BOW	100.0	55.0	51.3	50.9
SDAE	100.0	92.9	63.7	99.0
ST-UNI	100.0	93.2	62.3	96.6
ST-BI	96.6	88.7	63.2	74.7
InferSent	100.0	86.4	50.1	97.2

Classification accuracy

	CONTENT	ORDER	ROLE	NEG
BOW	100.0	55.0	51.3	50.9
SDAE	100.0	92.9	63.7	99.0
ST-UNI	100.0	93.2	62.3	96.6
ST-BI	96.6	88.7	63.2	74.7
InferSent	100.0	86.4	50.1	97.2

Classification accuracy

	CONTENT	ORDER	ROLE	NEG
BOW	100.0	55.0	51.3	50.9
SDAE	100.0	92.9	63.7	99.0
ST-UNI	100.0	93.2	62.3	96.6
ST-BI	96.6	88.7	63.2	74.7
InferSent	100.0	86.4	50.1	97.2

Classification accuracy

	CONTENT	ORDER	ROLE	NEG
BOW	100.0	55.0	51.3	50.9
SDAE	100.0	92.9	63.7	99.0
ST-UNI	100.0	93.2	62.3	96.6
ST-BI	96.6	88.7	63.2	74.7
InferSent	100.0	86.4	50.1	97.2

*The waitress is **not** actually **serving** the customer who is sleeping*

Classification accuracy

	CONTENT	ORDER	ROLE	NEG
BOW	100.0	55.0	51.3	50.9
SDAE	100.0	92.9	63.7	99.0
ST-UNI	100.0	93.2	62.3	96.6
ST-BI	96.6	88.7	63.2	74.7
InferSent	100.0	86.4	50.1	97.2

*The waitress is **not** actually **serving** the customer who is sleeping*

the waitress is not serving the customer	customer the serving not is waitress the

Classification accuracy

	CONTENT	ORDER	ROLE	NEG
BOW	100.0	55.0	51.3	50.9
SDAE	100.0	92.9	63.7	99.0
ST-UNI	100.0	93.2	62.3	96.6
ST-BI	96.6	88.7	63.2	74.7
InferSent	100.0	86.4	50.1	97.2

Classification accuracy

	CONTENT	ORDER	ROLE	NEG
BOW	100.0	55.0	51.3	50.9
SDAE	100.0	92.9	63.7	99.0
ST-UNI	100.0	93.2	62.3	96.6
ST-BI	96.6	88.7	63.2	74.7
InferSent	100.0	86.4	50.1	97.2

Classification accuracy

	CONTENT	ORDER	ROLE	NEG
BOW	100.0	55.0	51.3	50.9
InferSent	100.0	86.4	50.1	97.2

Sequence models appear to identify linking of negation to next verb

Work to be done on semantic roles

Update from Sesame Street

- Davis Yoshida (TTIC) tested semantic role tasks on ELMo, BERT, GPT
- Tested various configurations: CLS token, average of WordPiece tokens – below reports best performance

- ELMo (68.60%)
- BERT (63.00%)
- GPT (61.4%)

Outline

1. Assessing systematic composition in sentence encoders
2. **Simpler models as approximation of real-time predictive response**
3. Evaluating pre-trained LMs against human predictive responses

Beyond the endpoint

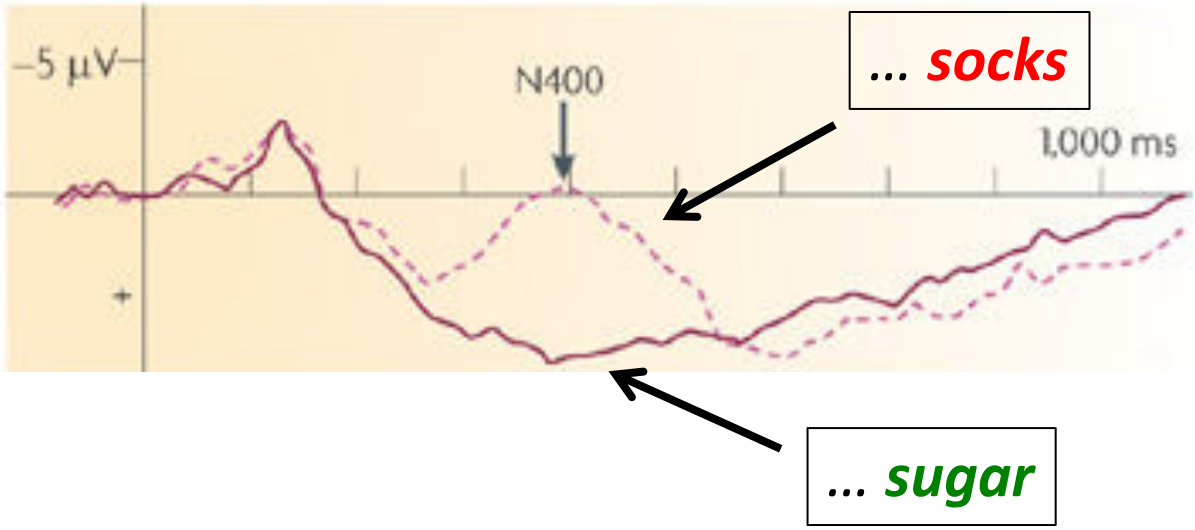
- Part I also emphasized that BOW can’t capture sentence meaning, so a model that resembles BOW can’t be doing understanding

- But there are other stages of comprehension that might actually look a bit like this

Measuring human brain activity (EEG)

N400 component

I take coffee with cream and _____



(Kutas & Hillyard, 1980)

Cloze probability

I take coffee with cream and _____

… ***socks*** Cloze probability = 0

… ***sugar*** Cloze probability = .6

Deviating from cloze

*The restaurant owner forgot which **customer** the **waitress** had ____*

Chow et al., 2015

Deviating from cloze

*The restaurant owner forgot which **customer** the **waitress** had ____*

*… **served***

Chow et al., 2015

Deviating from cloze

*The restaurant owner forgot which **customer** the **waitress** had ____*

*… **served***

*The restaurant owner forgot which **waitress** the **customer** had ____*

Chow et al., 2015

Deviating from cloze

*The restaurant owner forgot which **customer** the **waitress** had ____*

*… **served***

*The restaurant owner forgot which **waitress** the **customer** had ____*

*… **served***

Chow et al., 2015

Deviating from cloze

*The restaurant owner forgot which **customer** the **waitress** had ____*

*… **served***

*The restaurant owner forgot which **waitress** the **customer** had ____*

*… **served***

Chow et al., 2015

Deviating from cloze

*The restaurant owner forgot which **customer** the **waitress** had ____*

*… **served***

*The restaurant owner forgot which **waitress** the **customer** had ____*

*… **served***

Chow et al., 2015

N400

- Probably reflects most efficient available information for predicting upcoming words

- BOW-type representation may be a common go-to for this purpose

Federmeier & Kutas (1999)

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

Federmeier & Kutas (1999)

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

… football expected

Federmeier & Kutas (1999)

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

… *football*	expected
… *baseball*	within-category

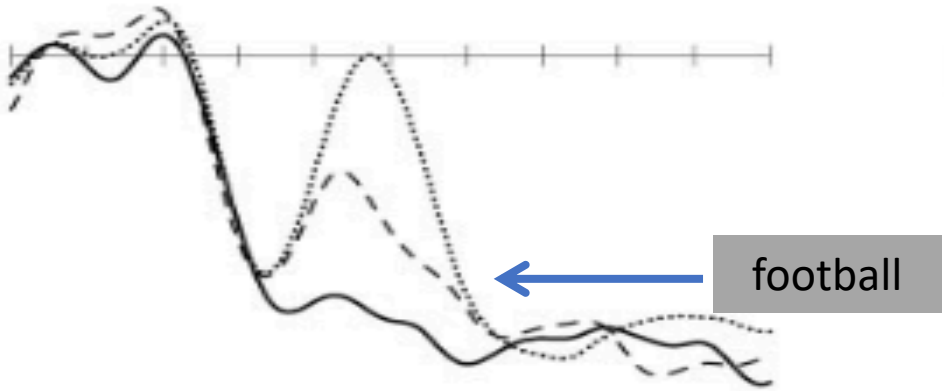
Federmeier & Kutas (1999)

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

… *football*	expected
… *baseball*	within-category
… *monopoly*	between-category

Federmeier & Kutas (1999)

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____



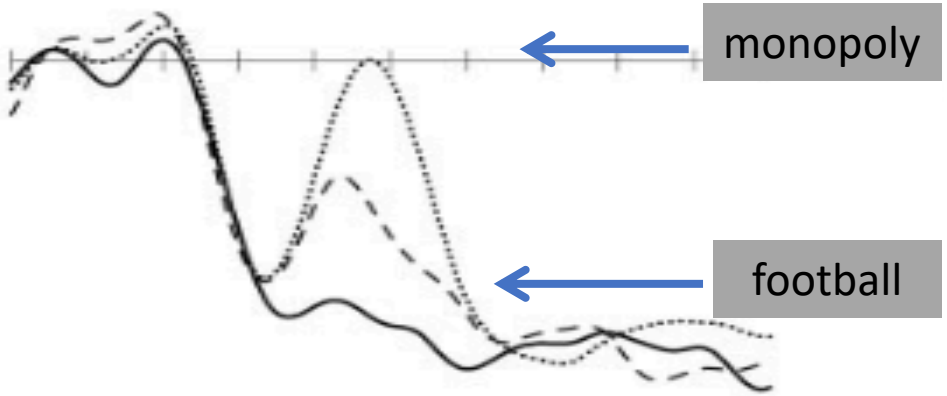
——— *Expected Exemplars*

- - - - - - *Within Category Violations*

..... *Between Category Violations*

Federmeier & Kutas (1999)

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____



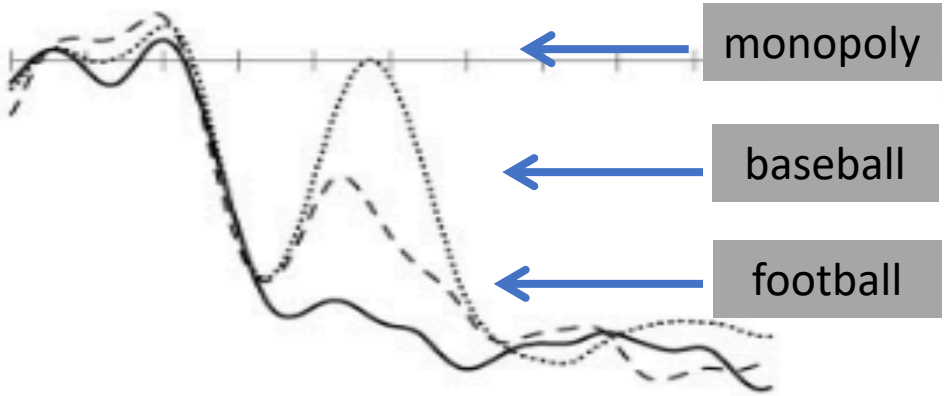
——— *Expected Exemplars*

- - - - - - *Within Category Violations*

..... *Between Category Violations*

Federmeier & Kutas (1999)

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____



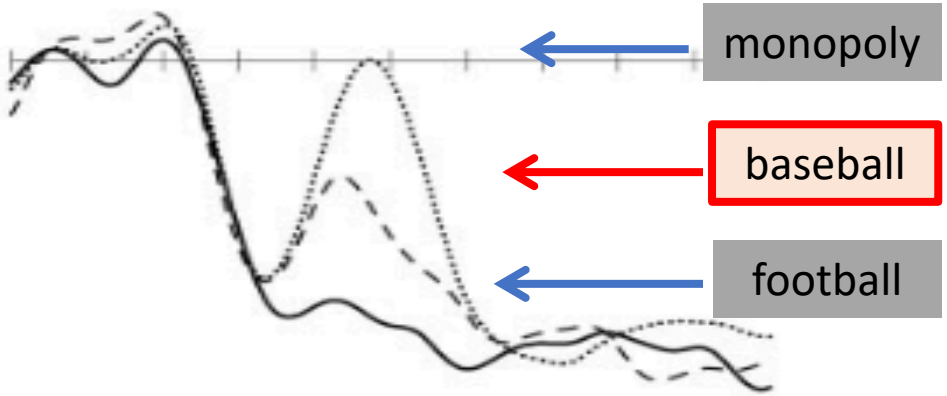
——— *Expected Exemplars*

- - - - - - *Within Category Violations*

..... *Between Category Violations*

Federmeier & Kutas (1999)

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____



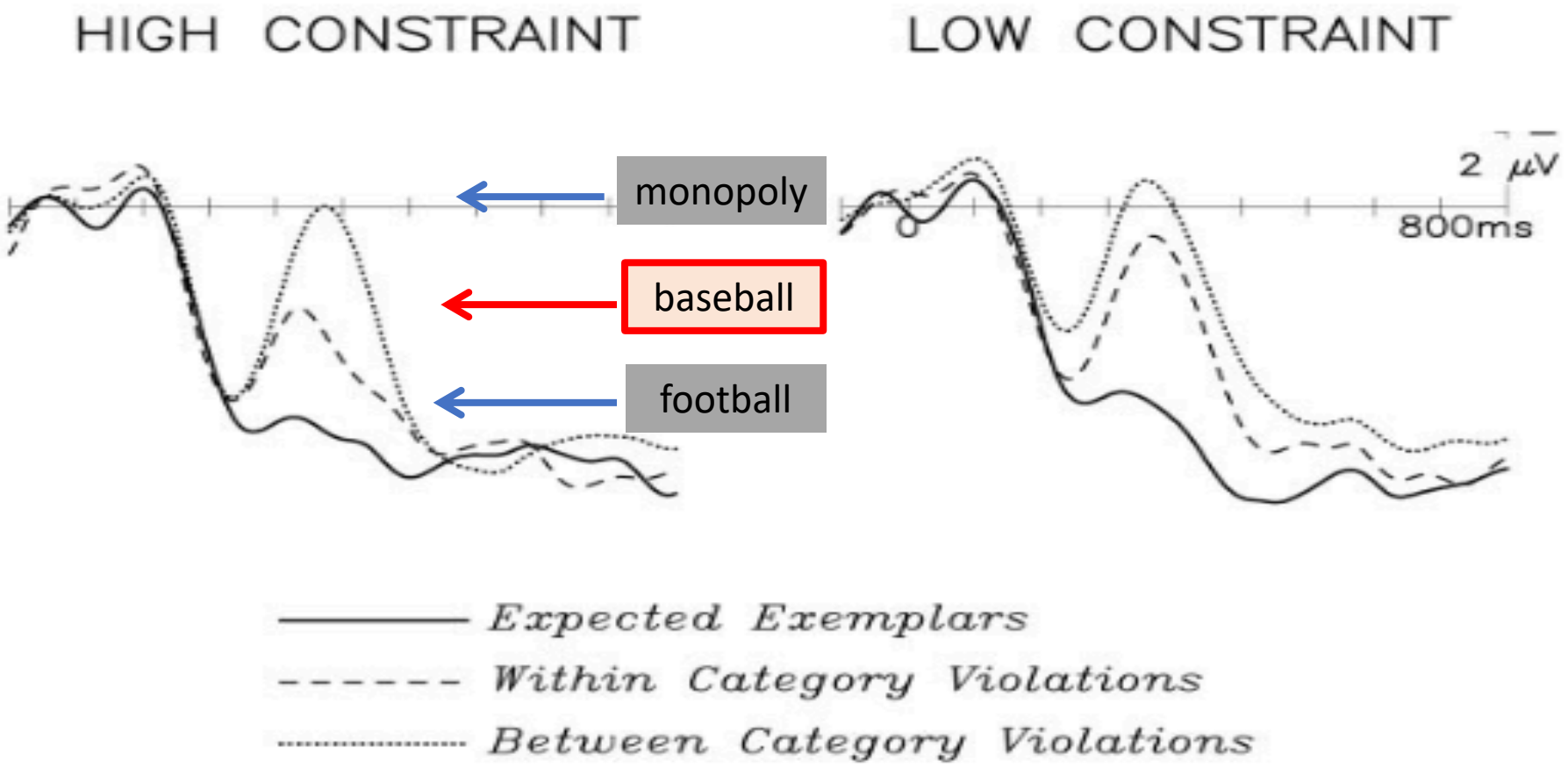
Unexpected facilitation

——— *Expected Exemplars*

- - - - - - *Within Category Violations*

..... *Between Category Violations*

Federmeier & Kutas (1999)

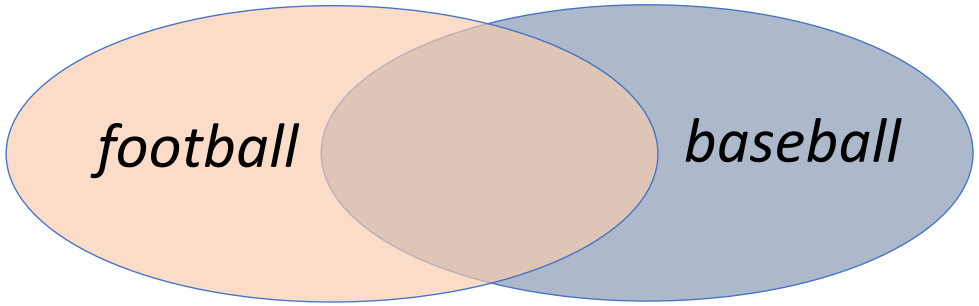


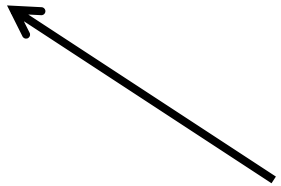
Federmeier & Kutas account

football

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

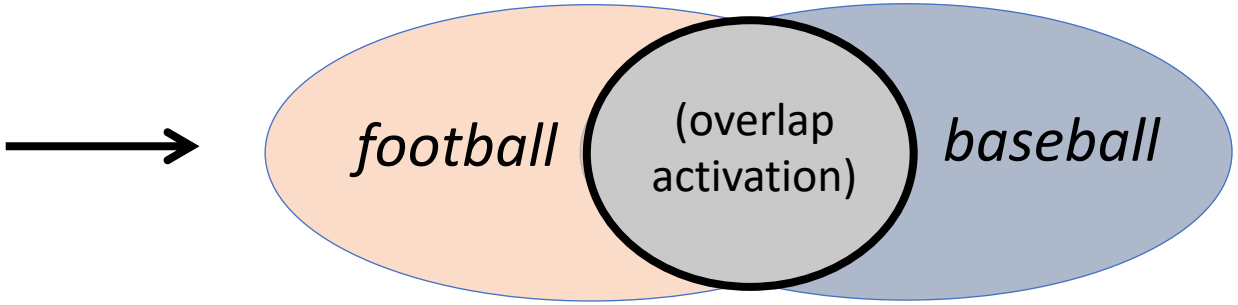
Federmeier & Kutas account





He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

Federmeier & Kutas account



He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

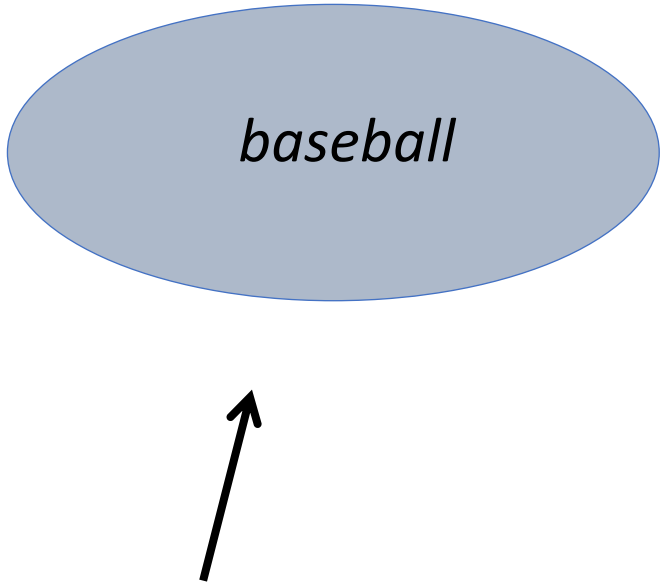
Alternative account

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

Alternative account

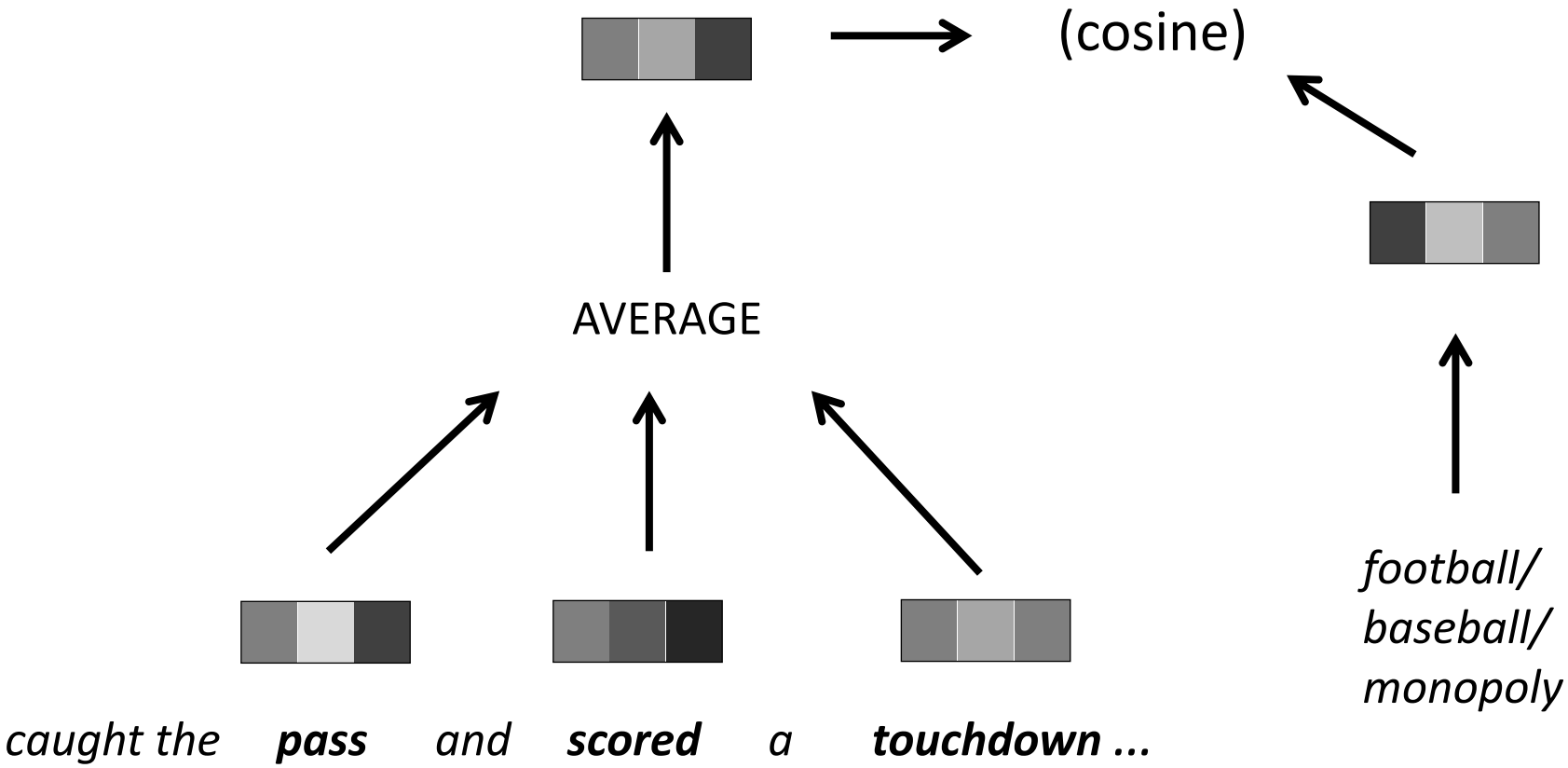
He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

Alternative account

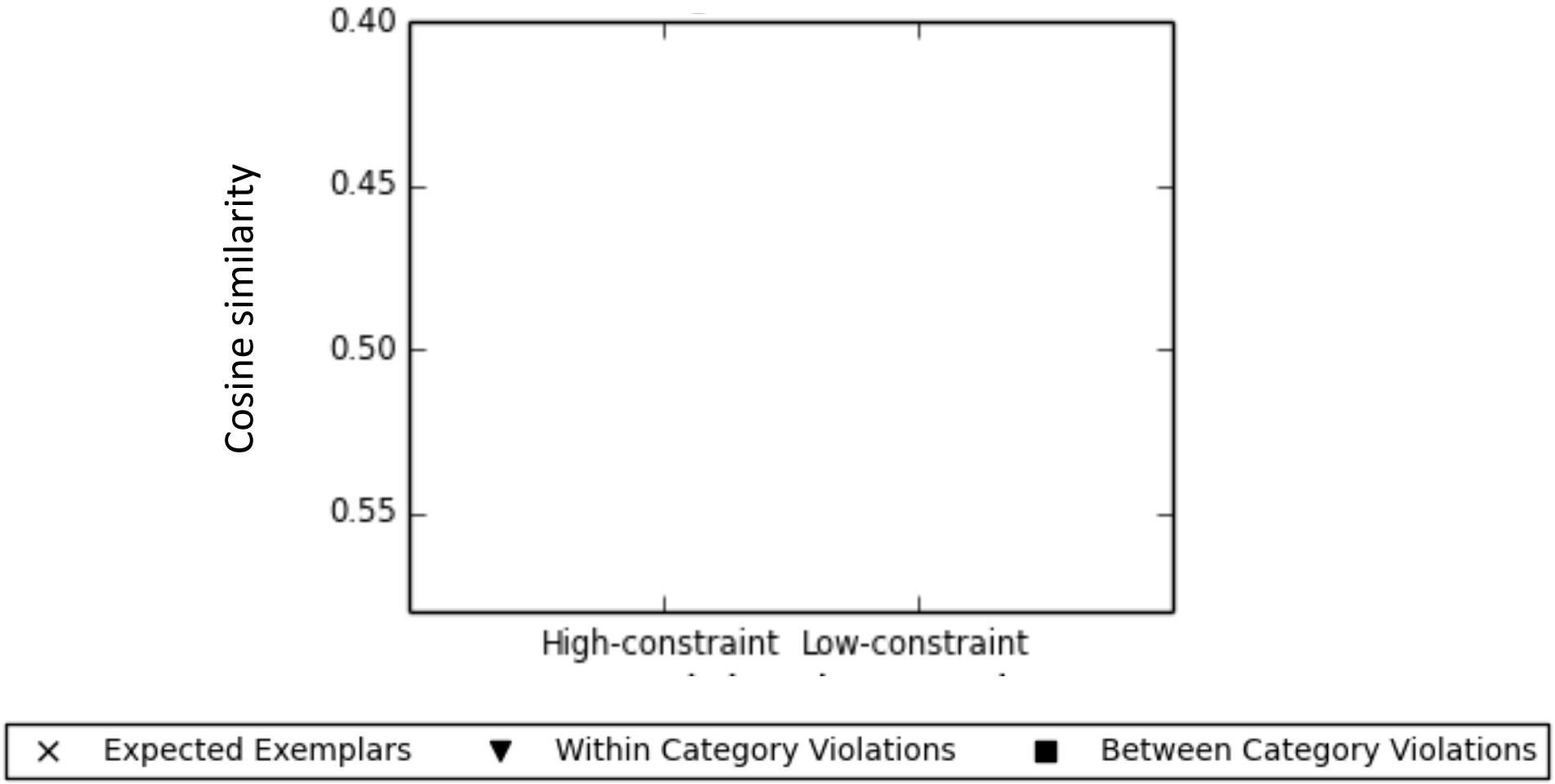


He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

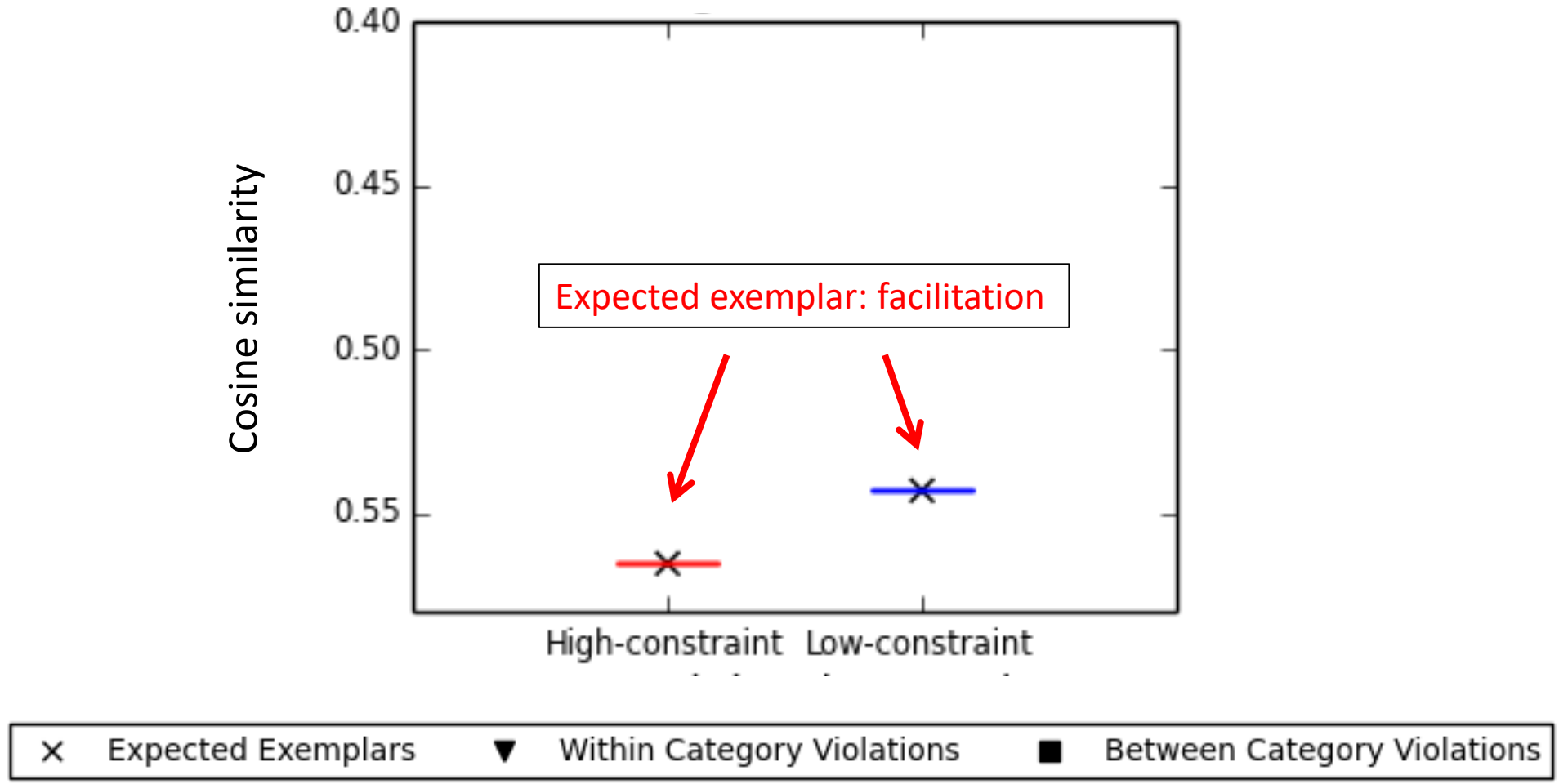
BOW averaging simulation



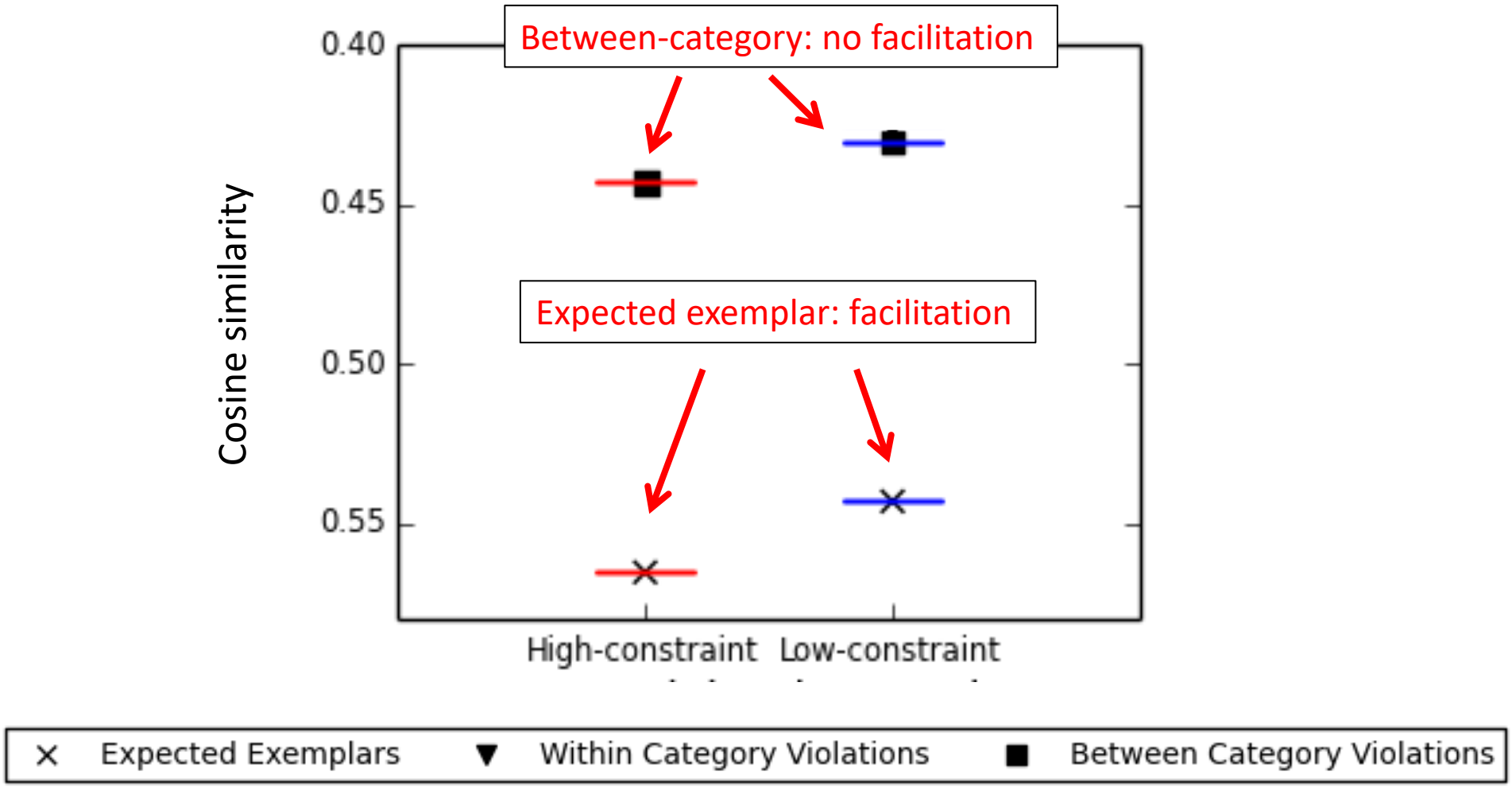
Simulation results



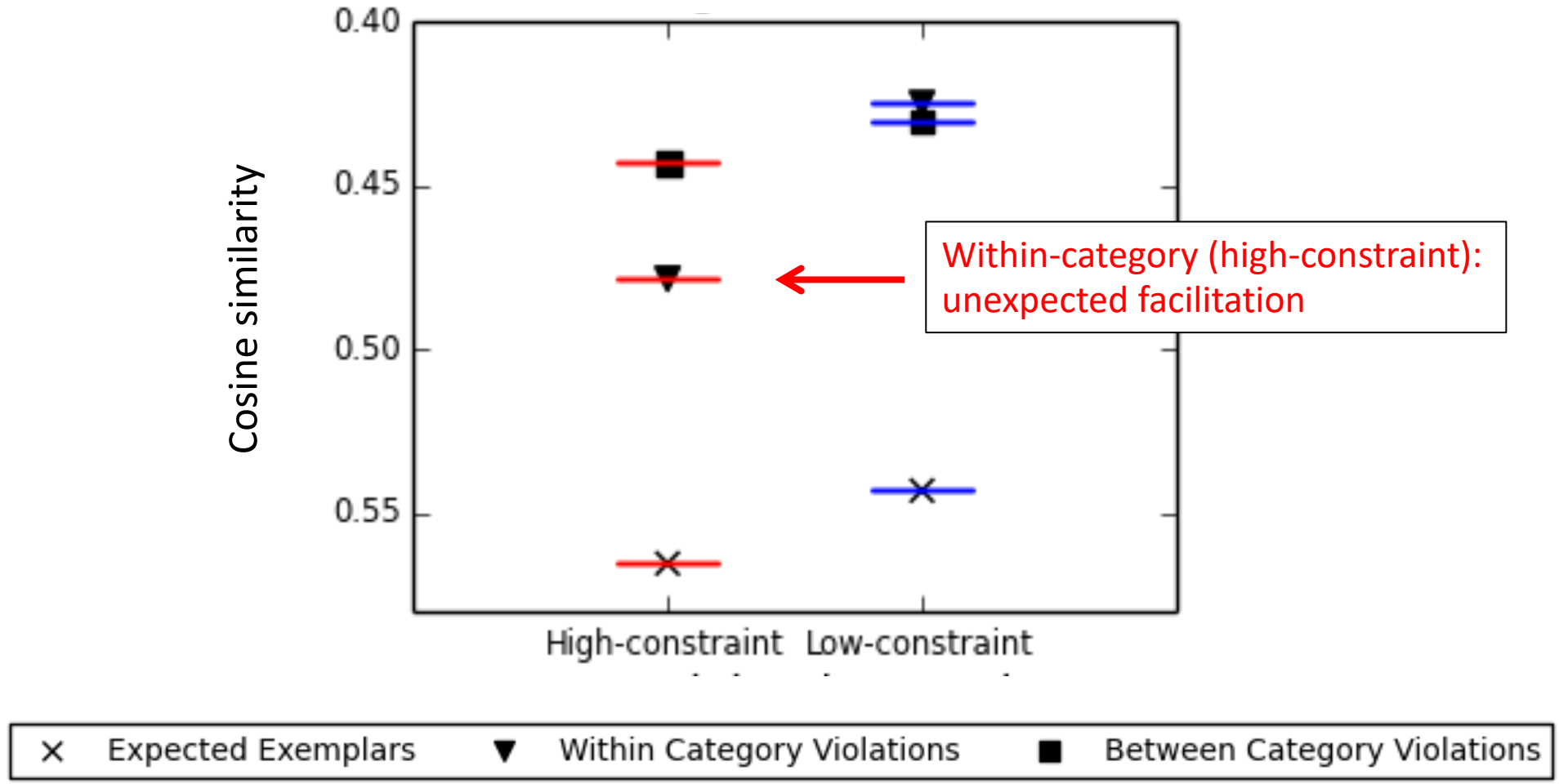
Simulation results



Simulation results



Simulation results



Interim takeaways

- Cognitive scientists: alternative explanation for observed result (made possible by availability of word embeddings)
- Our purposes: BOW model may not amount to comprehension – but it may align with other aspects of human processing
- Understanding which part of human processing we are approximating can help to improve in desired directions

Outline

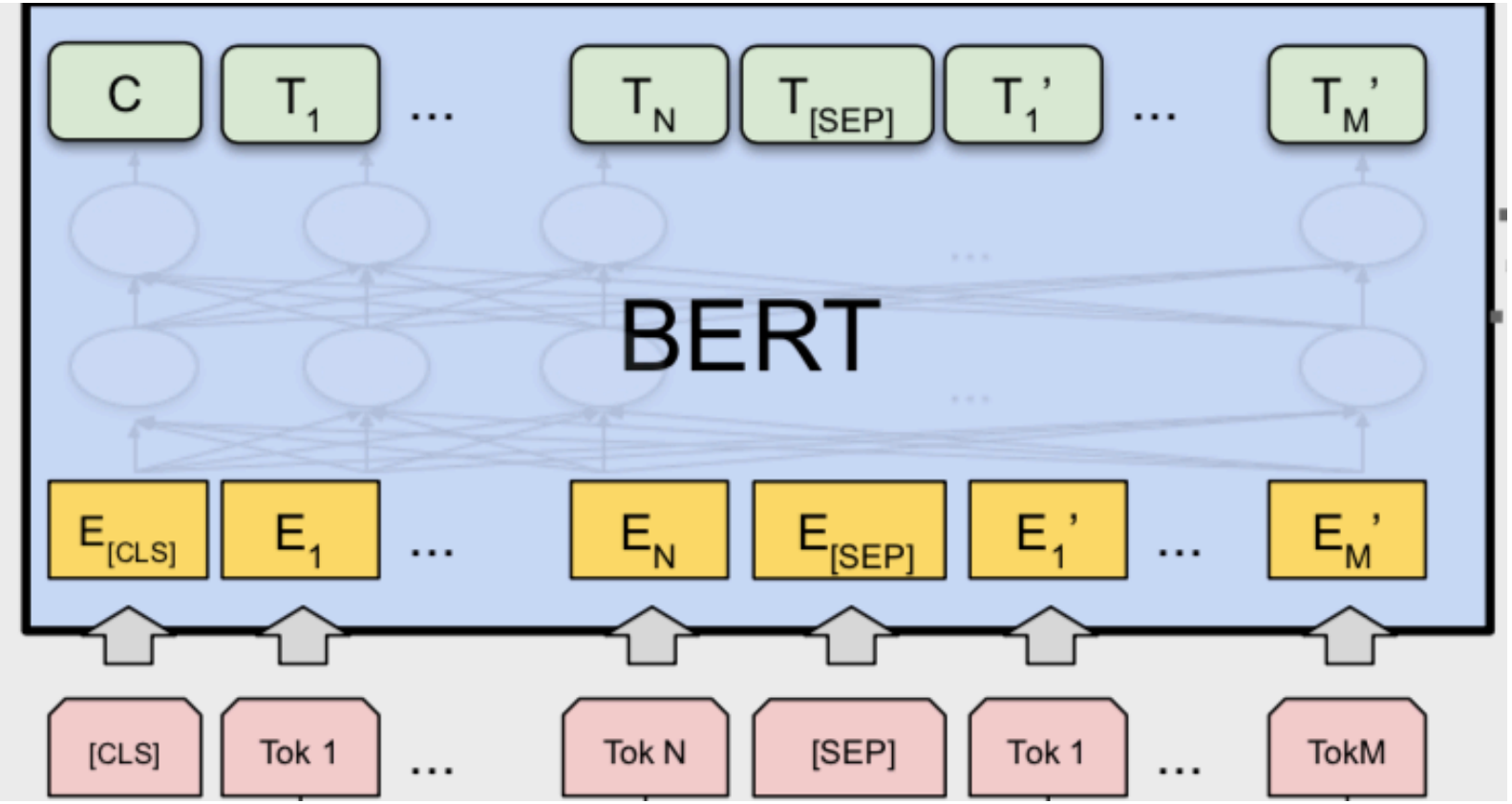
1. Assessing systematic composition in sentence encoders
2. Simpler models as approximation of real-time predictive response
3. **Evaluating pre-trained LMs against human predictive responses**

Pre-trained language models

- Impressive generalization across large number of tasks
- What kinds of generalizable linguistic competence do these models acquire during LM pre-training?
- Is it “understanding”? Is it shallower?

BERT

(Devlin et al 2018)



Probe representations?

- We could use probing tasks to probe the representations that pretrained models produces
- Few a priori expectations
- Should the CLS token represent all the sentence information? Should the average of token representations? At which layers?

Test word predictions

- Alternative: test pre-trained BERT in its most natural setting of predicting words in context

- What information is BERT sensitive to when making word predictions in context?

Psycholinguistic tests

- Designed to draw conclusions based on predictive responses in context
- Controlled to ask targeted questions about predictive mechanisms

N400/cloze divergence

- Choose psycholinguistic tests for which the N400 and cloze response diverge
- N400 predictive response shows apparent insensitivity to certain useful information for prediction
- Will BERT show similar insensitivities, or will it be able to make use of the higher-level predictive information that cloze reflects?

Psycholinguistic diagnostics

- Adapt three psycholinguistic datasets

- Three types of tests for each:

1. Word prediction accuracy—how well can the model use the relevant information to guide word predictions
2. Sensitivity tests—how well can the model distinguish between completions that the N400 has showed insensitivity on
3. Qualitative analysis—what do BERT's top predictions tell us about the information it has access to?

Datasets

- CPRAG-102: commonsense/pragmatic inference
- ROLE-88: event knowledge and semantic roles
- NEG-136: negation

Datasets

- **CPRAG-102: commonsense/pragmatic inference**
- ROLE-88: event knowledge and semantic roles
- NEG-136: negation

CPRAG-102: commonsense/pragmatic inference

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

He complained that after she kissed him, he couldn't get the red color off his face. He finally just asked her to stop wearing that ____

Prediction accuracy test

- Need to use commonsense inference to discern what is being described in first sentence

- Need to use pragmatic inference (along with normal syntactic/semantic information) to determine how the second sentence relates to the first

CPRAG-102: commonsense/pragmatic inference

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

He complained that after she kissed him, he couldn't get the red color off his face. He finally just asked her to stop wearing that ____

Sensitivity test

- Can BERT distinguish between completions with semantic features in common?

CPRAG-102: commonsense/pragmatic inference

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

… football

Federmeier & Kutas (1999)

CPRAG-102: commonsense/pragmatic inference

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

… football

… baseball

Federmeier & Kutas (1999)

CPRAG-102: commonsense/pragmatic inference

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of ____

… football

… baseball

… monopoly

Federmeier & Kutas (1999)

Datasets

- CPRAG-102: commonsense/pragmatic inference
- **ROLE-88: event knowledge and semantic roles**
- NEG-136: negation

ROLE-88: events and semantic roles

*The restaurant owner forgot which **customer** the **waitress** had ____*

*The restaurant owner forgot which **waitress** the **customer** had ____*

Original study: Chow et al., 2015

Prediction accuracy test

- Need to use semantic role information and knowledge about typical events in order to make accurate predictions

ROLE-88: events and semantic roles

*The restaurant owner forgot which **customer** the **waitress** had ____*

*The restaurant owner forgot which **waitress** the **customer** had ____*

Original study: Chow et al., 2015

Sensitivity test

- Will BERT reliably prefer continuations in the appropriate contexts rather than the role-reversed contexts?

ROLE-88: events and semantic roles

*The restaurant owner forgot which **customer** the **waitress** had ____*

*… **served***

Chow et al., 2015

ROLE-88: events and semantic roles

*The restaurant owner forgot which **customer** the **waitress** had ____*

*… **served***

*The restaurant owner forgot which **waitress** the **customer** had ____*

Chow et al., 2015

ROLE-88: events and semantic roles

*The restaurant owner forgot which **customer** the **waitress** had ____*

*… **served***

*The restaurant owner forgot which **waitress** the **customer** had ____*

*… **served***

Chow et al., 2015

Datasets

- CPRAG-102: commonsense/pragmatic inference
- ROLE-88: event knowledge and semantic roles
- **NEG-136: negation**

NEG-136: negation

A robin is a ____

Original study: Fischler et al., 1983

NEG-136: negation

A robin is a ____

*… **bird***

Original study: Fischler et al., 1983

NEG-136: negation

A robin is a ____

*… **bird***

A robin is not a ____

Original study: Fischler et al., 1983

NEG-136: negation

A robin is a ____

*… **bird***

A robin is not a ____

*… **bird***

Original study: Fischler et al., 1983

NEG-136: negation

A robin is a ____

*… **bird***

A robin is not a ____

*… **bird***

Original study: Fischler et al., 1983

Prediction accuracy

- This test doesn’t make sense in negated contexts, so test accuracy only on affirmative contexts

- Accurate predictions here require access to hypernym information

Sensitivity test

- This is where the test of negation comes in
- Can BERT prefer true continuations to false continuations, with and without negation?

Experiments

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Experiments

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of [MASK]

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Experiments

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of [MASK]

Extract BERT word predictions on [MASK] token, as in pre-training

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Experiments

- $BERT_{Base}$ – 12 layers, hidden layer size 768 dimensions, 12 self-attention heads. Total parameters 110M

- $BERT_{Large}$ – 24 layers, 1024 dim hidden size, 16 self-attention heads. Total parameters 340M

Results: CPRAG accuracy test

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of [MASK]

football in top k BERT predictions ?

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: CPRAG accuracy test

	Orig
$\text{BERT}_{\text{BASE}}\ k = 1$	23.5
$\text{BERT}_{\text{LARGE}}\ k = 1$	35.3
$\text{BERT}_{\text{BASE}}\ k = 5$	52.9
$\text{BERT}_{\text{LARGE}}\ k = 5$	52.9

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: CPRAG accuracy test

	Orig	Shuf	Trunc
$\text{BERT}_{\text{BASE}}\ k = 1$	23.5	14.1 ± 3.1	14.7
$\text{BERT}_{\text{LARGE}}\ k = 1$	35.3	17.4 ± 3.5	17.6
$\text{BERT}_{\text{BASE}}\ k = 5$	52.9	36.1 ± 2.8	35.3
$\text{BERT}_{\text{LARGE}}\ k = 5$	52.9	39.2 ± 3.9	32.4

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: CPRAG accuracy test

	Orig	Shuf	Trunc	Shuf + Trunc
$\text{BERT}_{\text{BASE}}\ k = 1$	23.5	14.1 ± 3.1	14.7	8.1 ± 3.4
$\text{BERT}_{\text{LARGE}}\ k = 1$	35.3	17.4 ± 3.5	17.6	10.0 ± 3.0
$\text{BERT}_{\text{BASE}}\ k = 5$	52.9	36.1 ± 2.8	35.3	22.1 ± 3.2
$\text{BERT}_{\text{LARGE}}\ k = 5$	52.9	39.2 ± 3.9	32.4	21.3 ± 3.7

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: CPRAG sensitivity test

He caught the pass and scored another touchdown. There was nothing he enjoyed more than a good game of [MASK]

football >

baseball and *monopoly* ?

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: CPRAG sensitivity test

	Prefer good	w/ .01 thresh
$\text{BERT}_{\text{BASE}}$	73.5	44.1
$\text{BERT}_{\text{LARGE}}$	79.4	58.8

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

CPRAG qualitative analysis

Context	$\text{BERT}_{\text{LARGE}}$ predictions
Pablo wanted to cut the lumber he had bought to make some shelves. He asked his neighbor if he could borrow her ____	*car, house, room, truck, apartment*
The snow had piled up on the drive so high that they couldn't get the car out. When Albert woke up, his father handed him a ____	*note, letter, gun, blanket, newspaper*
At the zoo, my sister asked if they painted the black and white stripes on the animal. I explained to her that they were natural features of a ____	*cat, person, human, bird, species*

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: ROLE accuracy test

*The restaurant owner forgot which **customer** the **waitress** had [MASK]*

(*served* in top k BERT predictions?)

*The restaurant owner forgot which **waitress** the **customer** had [MASK]*

(*tipped* in top k BERT predictions?)

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: ROLE accuracy test

	Orig
$\text{BERT}_{\text{BASE}}$ k=1	14.8
$\text{BERT}_{\text{LARGE}}$ k=1	13.6
$\text{BERT}_{\text{BASE}}$ k=5	27.3
$\text{BERT}_{\text{LARGE}}$ k=5	37.5

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: ROLE accuracy test

	Orig	-Obj	-Sub
$\text{BERT}_{\text{BASE}}$ k=1	14.8	12.5	12.5
$\text{BERT}_{\text{LARGE}}$ k=1	13.6	5.7	6.8
$\text{BERT}_{\text{BASE}}$ k=5	27.3	26.1	22.7
$\text{BERT}_{\text{LARGE}}$ k=5	37.5	18.2	21.6

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: ROLE accuracy test

	Orig	-Obj	-Sub	-Both
$\text{BERT}_{\text{BASE}}$ k=1	14.8	12.5	12.5	9.1
$\text{BERT}_{\text{LARGE}}$ k=1	13.6	5.7	6.8	4.5
$\text{BERT}_{\text{BASE}}$ k=5	27.3	26.1	22.7	18.2
$\text{BERT}_{\text{LARGE}}$ k=5	37.5	18.2	21.6	14.8

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: ROLE sensitivity test

*The restaurant owner forgot which **customer** the **waitress** had [MASK]*

served

>

*The restaurant owner forgot which **waitress** the **customer** had [MASK]*

served

?

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: ROLE sensitivity test

	Prefer good	w/ .01 thresh
$\text{BERT}_{\text{BASE}}$	75.0	31.8
$\text{BERT}_{\text{LARGE}}$	86.4	43.2

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

ROLE qualitative analysis

Context	$\text{BERT}_{\text{BASE}}$ predictions	$\text{BERT}_{\text{LARGE}}$ predictions
the camper reported which girl the bear had ____	*taken, killed, attacked, bitten, picked*	*attacked, killed, eaten, taken, targeted*
the camper reported which bear the girl had ____	*taken, killed, fallen, bitten, jumped*	*taken, left, entered, found, chosen*
the restaurant owner forgot which customer the waitress had ____	*served, hired, brought, been, taken*	*served, been, delivered, mentioned, brought*
the restaurant owner forgot which waitress the customer had ____	*served, been, chosen, ordered, hired*	*served, chosen, called, ordered, been*

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: NEG accuracy test

A robin is a [MASK]

bird in top k BERT predictions ?

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: NEG accuracy test

	Accuracy
$\text{BERT}_{\text{BASE}}$ $k = 1$	38.9
$\text{BERT}_{\text{LARGE}}$ $k = 1$	44.4
$\text{BERT}_{\text{BASE}}$ $k = 5$	100
$\text{BERT}_{\text{LARGE}}$ $k = 5$	100

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: NEG sensitivity test

A robin is a [MASK]

bird > *tree* ?

A robin is not a [MASK]

tree > *bird* ?

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Results: NEG sensitivity test

	Affirmative	Negative
$\text{BERT}_{\text{BASE}}$	100	0.0
$\text{BERT}_{\text{LARGE}}$	100	0.0

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

NEG qualitative analysis

Context	$\text{BERT}_{\text{LARGE}}$ predictions
A robin is a ____	*bird, robin, person, hunter, pigeon*
A daisy is a ____	*daisy, rose, flower, berry, tree*
A hammer is a ____	*hammer, tool, weapon, nail, device*
A hammer is an ____	*object, instrument, axe, implement, explosive*
A robin is not a ____	*robin, bird, penguin, man, fly*
A daisy is not a ____	*daisy, rose, flower, lily, cherry*
A hammer is not a ____	*hammer, weapon, tool, gun, rock*
A hammer is not an ____	*object, instrument, axe, animal, artifact*

Ettinger (2019). *What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models*

Takeaways

- Decent on sensitivity to role reversal and differences within semantic category – but seemingly weaker sensitivity than cloze
- Great with hypernyms, determiners, grammaticality
- Struggles with challenging inference and event-based prediction
- Clear insensitivity to contextual impacts of negation

Discussion

- Many of these results give general indication that these pre-trained models have a way to go to incorporate human inference
- Negation result is more striking and starker
- Not surprising, ultimately, given LM training – but possibly means that LM training isn’t suited for learning negation
- What other aspects of comprehension have this property?

Outline

1. Assessing systematic composition in sentence encoders
2. Simpler models as approximation of real-time predictive response
3. Evaluating pre-trained LMs against human predictive responses

Conclusions

- What we want to be able to do is capture the endpoint of comprehension
- What we're good at right now is leveraging co-occurrence statistics in a way that maximizes our ability to predict surrounding/upcoming words
- This sometimes causes our models to better resemble earlier stages of human comprehension rather than the endpoint
- Understanding what part of human processing we're capturing, and how that relates to what we do want to capture, could help us meet our goals

Thank you!

Philip Resnik

Colin Phillips

Naomi Feldman

GRF Grant DGE-1322106

NRT Grant DGE-1449815

Toyota Technological Institute at Chicago

Ahmed Elgohary