Principles and Interactive Tools for Evaluating and Improving the Behavior of NLP models

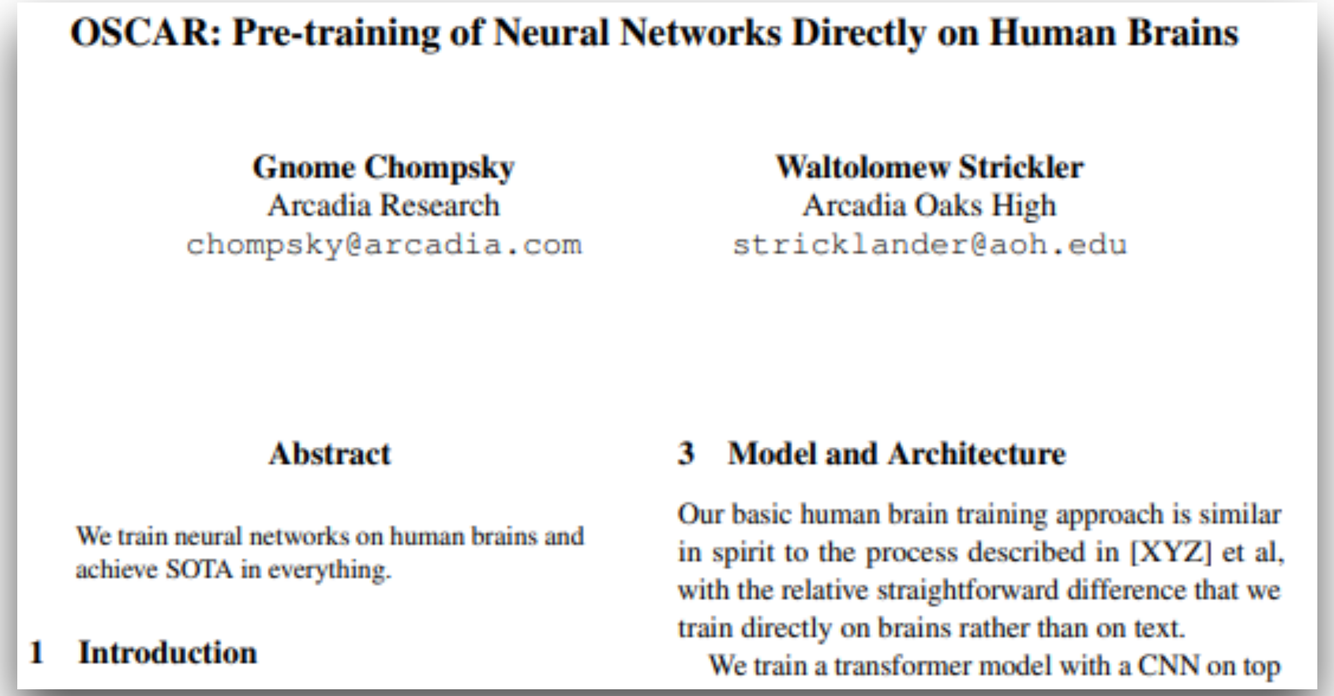
Tongshuang (Sherry) Wu
@tongshuangwu / wtshuang@cs.washington.edu
University of Washington

How do I check if my model works?

Should I replace my doctor with OSCAR?

Should we use OSCAR in our products?

If not, what do I need to fix?



OSCAR: Pre-training of Neural Networks Directly on Human Brains

Gnome Chompsky
Arcadia Research
chompsky@arcadia.com

Waltolomew Strickler
Arcadia Oaks High
stricklander@aoh.edu

Abstract

We train neural networks on human brains and achieve SOTA in everything.

1 Introduction

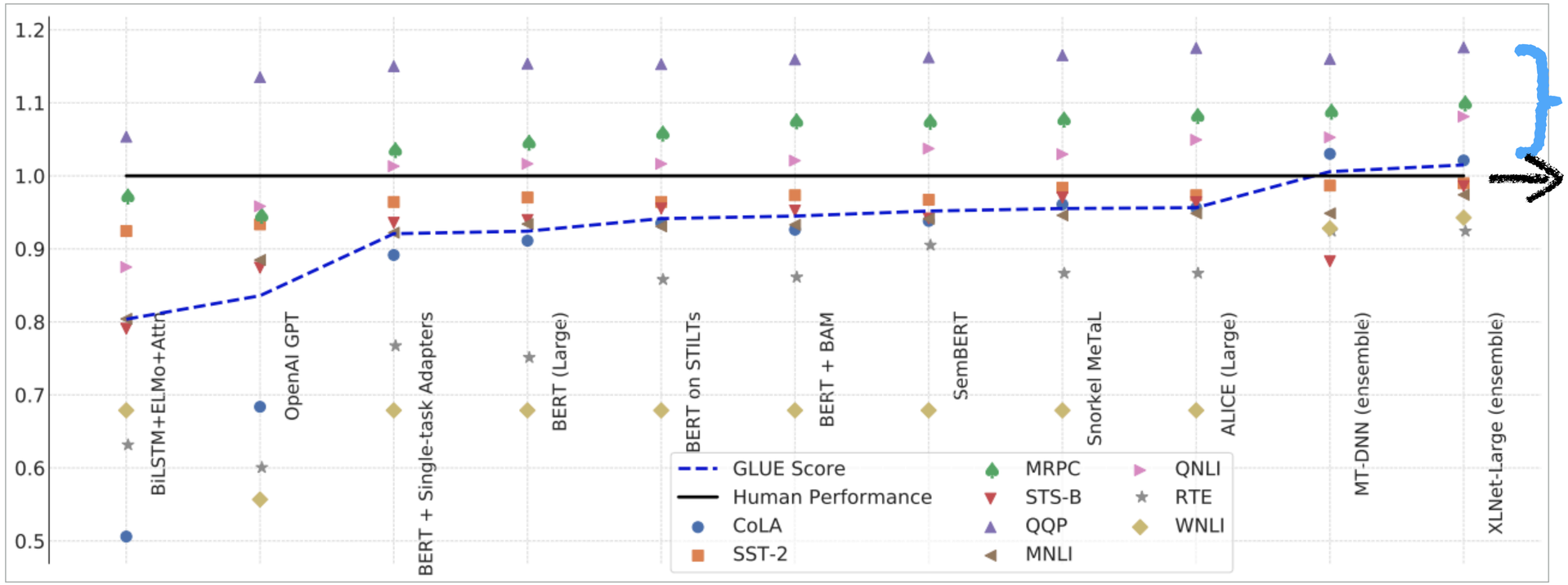
3 Model and Architecture

Our basic human brain training approach is similar in spirit to the process described in [XYZ] et al, with the relative straightforward difference that we train directly on brains rather than on text.

We train a transformer model with a CNN on top

Accuracy seems a good solution?

GLUE: "performance on the benchmark has recently come close to the level of non-expert humans, suggesting limited headroom for further research."



What could go wrong?

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... & Bowman, S. (2019). Superglue: A stickier benchmark for general-purpose language understanding systems. In Advances in Neural Information Processing Systems (pp. 3266-3280).

Missing critical data slices (bias, fairness)

What's in the figure?

```
> ceremony, wedding, bride, groom, dress
```

```
> person, people
```

https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

High accuracy ≠ Model succeeding.

How can we make sure our model can handle particular data slices?

Shortcuts/right for wrong reasons

What is the moustache made of?

> Banana

What are the eyes made of?

> Banana

What is?

> Banana

What?

> Banana

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models. arXiv preprint arXiv:1606.07356.

High accuracy ≠ Model succeeding.

How can we make sure our model can handle particular data slices?

Correct prediction ≠ correct reasoning.

How can practitioners ensure the model learns important features & avoid spurious correlations?

Analyzing structured data is easy.

How good is our model on records with different city entries?

What happens if I change the city column to ***`New York`****?*

Give me 0.001 seconds to run a SQL script!

Analyzing ~~structured data~~ text is ~~easy~~ hard.

How good is our model on passive sentences on cities?

What happens if I change the passive voice to positive?

Ugh…POS? Named entities? clustering?

“State-of-the-art”

we sampled 200 question answer pairs and manually analyzed their properties.

Joshi et al. ACL’17

Chen et al. ACL’16

We randomly select 50 incorrect questions and categorize them into 6 classes.

We sample 100 incorrect predictions and try to find common error categories.

Wadhwa et al. ACL’18

Joshi, Mandar, et al. "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension." arXiv preprint arXiv:1705.03551 (2017).
Chen, Danqi, Jason Bolton, and Christopher D. Manning. "A thorough examination of the cnn/daily mail reading comprehension task." arXiv preprint arXiv:1606.02858 (2016)
.Wadhwa, Soumya, Khyathi Raghavi Chandu, and Eric Nyberg. "Comparative analysis of neural qa models on squad." arXiv preprint arXiv:1806.06972 (2018).

“State-of-the-art”

we sampled 200 question answer pairs and manually analyze

Joshi et al.
ACL’17

Chen et a
ACL’16

them into 6 classes.

orize

We sar
comm

dhwa et al.
ACL’18

“We randomly select 50-100 instances and roughly label them into N error groups.”

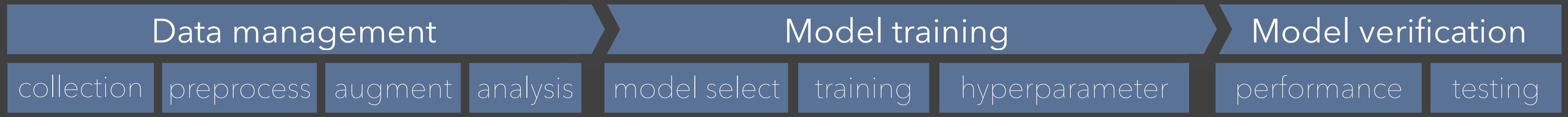
Under-representative, subjective, high variance, low reproducibility
(TOCHI 19, ACL 19, CHI 21)

Joshi, Mandar, et al. "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension." arXiv preprint arXiv:1705.03551 (2017).
Chen, Danqi, Jason Bolton, and Christopher D. Manning. "A thorough examination of the cnn/daily mail reading comprehension task." arXiv preprint arXiv:1606.02858 (2016)
.Wadhwa, Soumya, Khyathi Raghavi Chandu, and Eric Nyberg. "Comparative analysis of neural qa models on squad." arXiv preprint arXiv:1806.06972 (2018).

I **uncover pitfalls in the status-quo analysis** process, and

help NLP practitioners **inspect the inputs and outputs** of their models, such that they can gain **more systematic insights** into their models' behaviors.

I will focus on this in the talk!



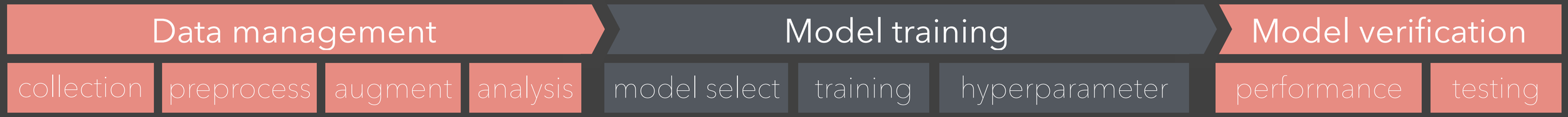
(Paleyes et al. 2020)

Paleyes, A., Urma, R. G., & Lawrence, N. D. (2020). Challenges in Deploying Machine Learning: a Survey of Case Studies. arXiv

I **uncover pitfalls in the status-quo analysis** process, and

help NLP practitioners **inspect the inputs and outputs** of their models, such that they can gain **more systematic insights** into their models' behaviors.

I will focus on this in the talk!



To improve

(Paleyes et al. 2020)

Paleyes, A., Urma, R. G., & Lawrence, N. D. (2020). Challenges in Deploying Machine Learning: a Survey of Case Studies. arXiv

Building blocks

High accuracy ≠ Model succeeding.

How can we make sure our model can handle particular data slices?

Correct prediction ≠ correct reasoning.

How can practitioners ensure the model learns important features & avoid superficial correlations?

Building blocks

Quantitative grouping

Inspect similar instances, semantically & syntactically

```
ENT(g) != ""
  and count(token(c, pattern=ENT(g))) >
    count(token(g, pattern=ENT(g)))
  and ENT(g) == ENT(p(m))
  and f1(m) == 0
```

Correct prediction ≠ correct reasoning.

How can practitioners ensure the model learns important features & avoid superficial correlations?

Building blocks

Quantitative grouping

Inspect similar instances, semantically & syntactically

```
ENT(g) != ""
  and count(token(c, pattern=ENT(g))) >
    count(token(g, pattern=ENT(g)))
  and ENT(g) == ENT(p(m))
  and f1(m) == 0
```

Counterfactual perturbation

Isolate important components targeted minimal rewrites

This is a good movie.

Lexical

This is ~~good~~ great a movie.

This is a good ~~movie~~ film.

Negation

This would be a good movie.

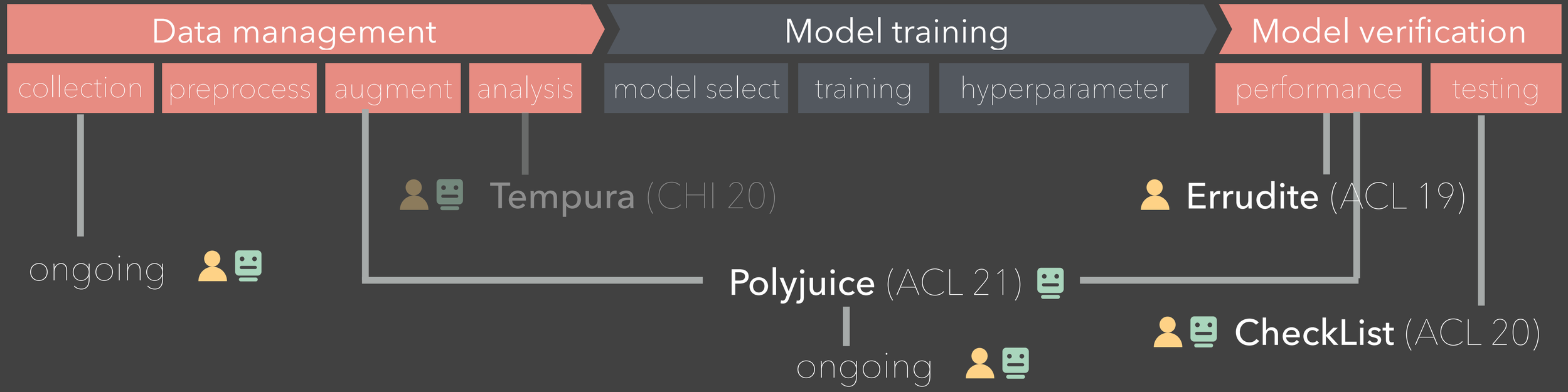
This is not a good movie.

Building blocks

Quantitative grouping

Counterfactual perturbation

Model development stages (Paleyes et al. 2020)



Model verification

performance

ACL 2019

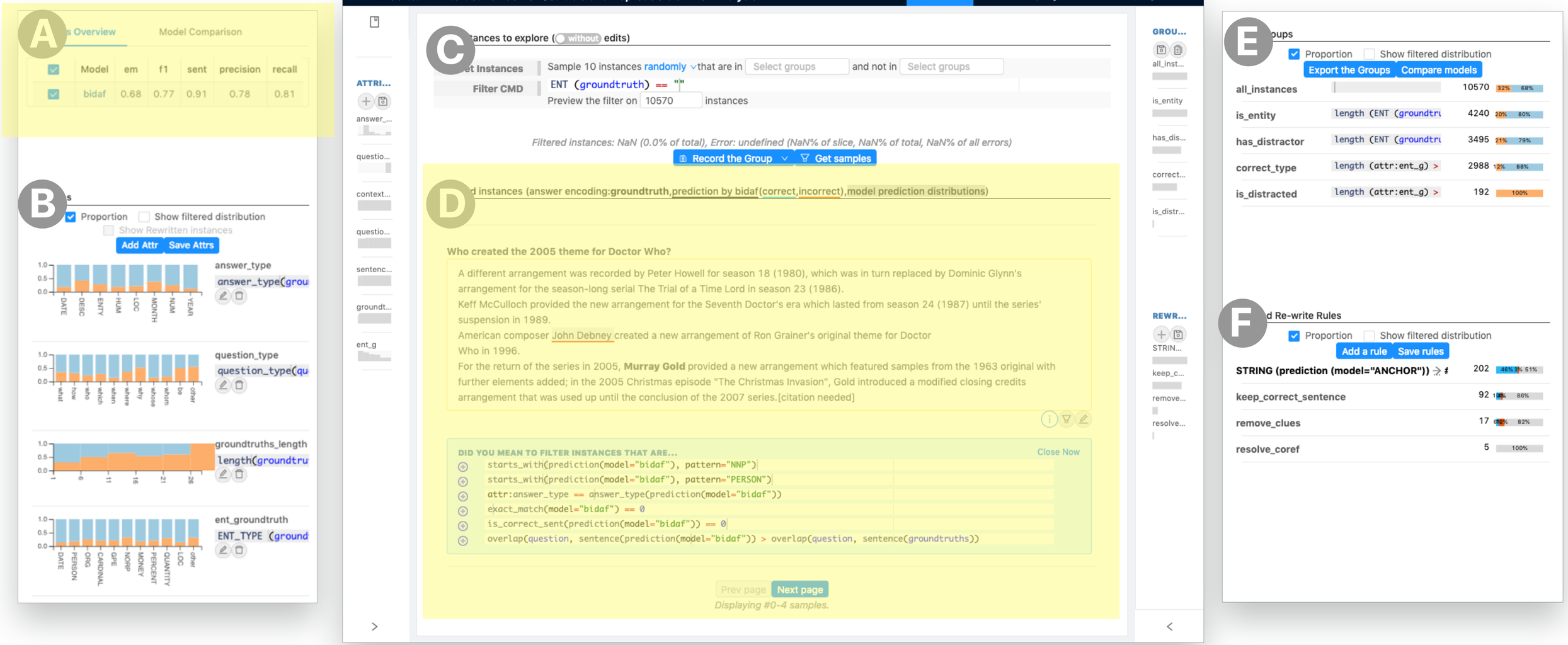
Errudite: Scalable, Reproducible, and Testable Error Analysis

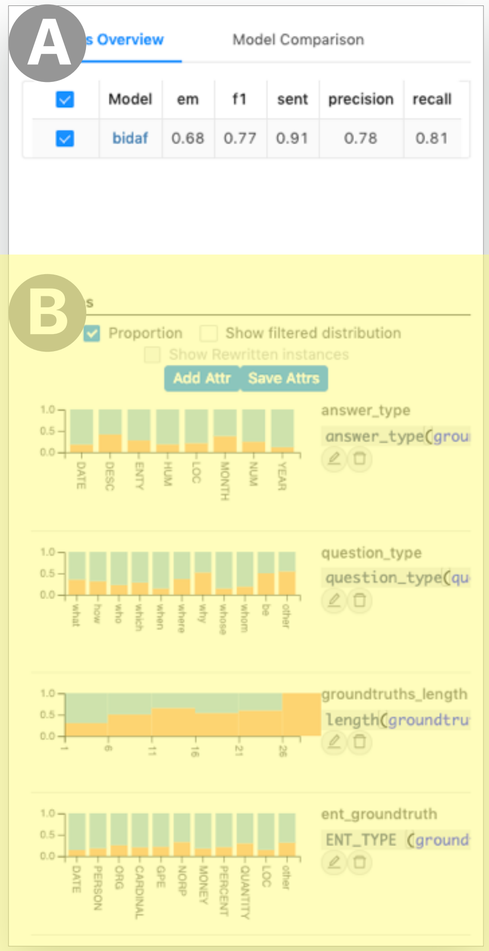
Tongshuang (Sherry) Wu @tongshuangwu
University of Washington

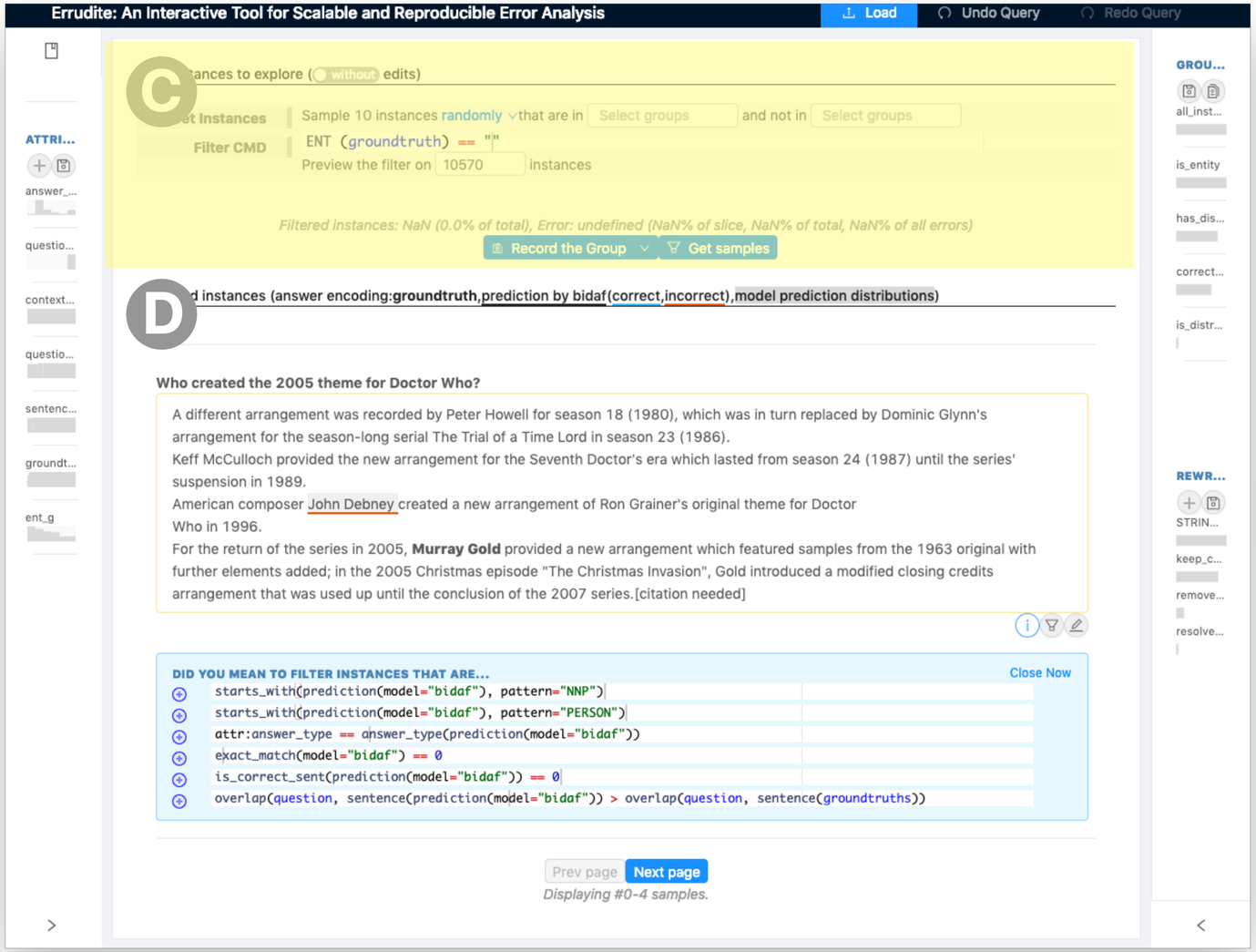
Marco Tulio Ribeiro
Microsoft Research

Jeffrey Heer @jeffrey_heer
Daniel S. Weld @dsweld
University of Washington





A
s Overview
Model Comparison
Model em f1 sent precision recall
bidaf 0.68 0.77 0.91 0.78 0.81
B
Proportion
Show filtered distribution
Show Rewritten instances
Add Attr
Save Attrs
answer_type
answer_type(grou
DATE DESC ENTY HUM LOC MONTH NUM YEAR
question_type
question_type(qu
what how who which when where why whose whom be other
groundtruths_length
length(groundtru
ent_groundtruth
ENT_TYPE (ground
DATE PERSON ORG CARDINAL GPE NORP MONEY PERCENT QUANTITY LOC other
Errudite: An Interactive Tool for Scalable and Reproducible Error Analysis
Load
Undo Query
Redo Query
C
ances to explore (without edits)
t Instances
Sample 10 instances randomly that are in
Select groups
and not in
Select groups
Filter CMD
ENT (groundtruth) == ""
Preview the filter on
10570
instances
Filtered instances: NaN (0.0% of total), Error: undefined (NaN% of slice, NaN% of total, NaN% of all errors)
Record the Group
Get samples
ATTRI...
answer_...
questio...
context...
questio...
sentenc...
groundt...
ent_g
GROU...
all_inst...
is_entity
has_dis...
correct...
is_distr...
REWR...
STRIN...
keep_c...
remove...
resolve...
D
d instances (answer encoding:groundtruth,prediction by bidaf(correct,incorrect),model prediction distributions)
Who created the 2005 theme for Doctor Who?
A different arrangement was recorded by Peter Howell for season 18 (1980), which was in turn replaced by Dominic Glynn's arrangement for the season-long serial The Trial of a Time Lord in season 23 (1986).
Keff McCulloch provided the new arrangement for the Seventh Doctor's era which lasted from season 24 (1987) until the series' suspension in 1989.
American composer John Debney created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996.
For the return of the series in 2005, Murray Gold provided a new arrangement which featured samples from the 1963 original with further elements added; in the 2005 Christmas episode "The Christmas Invasion", Gold introduced a modified closing credits arrangement that was used up until the conclusion of the 2007 series.[citation needed]
DID YOU MEAN TO FILTER INSTANCES THAT ARE...
Close Now
starts_with(prediction(model="bidaf"), pattern="NNP")
starts_with(prediction(model="bidaf"), pattern="PERSON")
attr:answer_type == answer_type(prediction(model="bidaf"))
exact_match(model="bidaf") == 0
is_correct_sent(prediction(model="bidaf")) == 0
overlap(question, sentence(prediction(model="bidaf")) > overlap(question, sentence(groundtruths))
Prev page
Next page
Displaying #0-4 samples.
E
oups
Proportion
Show filtered distribution
Export the Groups
Compare models
all_instances 10570
is_entity length (ENT (groundtru 4240
has_distractor length (ENT (groundtru 3495
correct_type length (attr:ent_g) > 2988
is_distracted length (attr:ent_g) > 192
F
d Re-write Rules
Proportion
Show filtered distribution
Add a rule
Save rules
STRING (prediction (model="ANCHOR")) ≠ 202
keep_correct_sentence 92
remove_clues 17
resolve_coref 5


A
s Overview
Model Comparison
Model em f1 sent precision recall
bidaf 0.68 0.77 0.91 0.78 0.81
B
Proportion
Show filtered distribution
Show Rewritten instances
Add Attr
Save Attrs
answer_type
answer_type(grou
DATE DESC ENTY HUM LOC MONTH NUM YEAR
question_type
question_type(qu
what how who which when where why whose whom be other
groundtruths_length
length(groundtru
ent_groundtruth
ENT_TYPE (ground
DATE PERSON ORG CARDINAL GPE NORP MONEY PERCENT QUANTITY LOC other


Errudite: An Interactive Tool for Scalable and Reproducible Error Analysis
Load
Undo Query
Redo Query
ATTRI...
answer_...
questio...
context...
questio...
sentenc...
groundt...
ent_g
C
ances to explore (without edits)
t Instances
Sample 10 instances randomly that are in
Select groups
and not in
Select groups
Filter CMD
ENT (groundtruth) == ""
Preview the filter on
10570
instances
Filtered instances: NaN (0.0% of total), Error: undefined (NaN% of slice, NaN% of total, NaN% of all errors)
Record the Group
Get samples
D
d instances (answer encoding:groundtruth,prediction by bidaf(correct,incorrect),model prediction distributions)
Who created the 2005 theme for Doctor Who?
A different arrangement was recorded by Peter Howell for season 18 (1980), which was in turn replaced by Dominic Glynn's arrangement for the season-long serial The Trial of a Time Lord in season 23 (1986).
Keff McCulloch provided the new arrangement for the Seventh Doctor's era which lasted from season 24 (1987) until the series' suspension in 1989.
American composer John Debney created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996.
For the return of the series in 2005, Murray Gold provided a new arrangement which featured samples from the 1963 original with further elements added; in the 2005 Christmas episode "The Christmas Invasion", Gold introduced a modified closing credits arrangement that was used up until the conclusion of the 2007 series.[citation needed]
DID YOU MEAN TO FILTER INSTANCES THAT ARE...
Close Now
starts_with(prediction(model="bidaf"), pattern="NNP")
starts_with(prediction(model="bidaf"), pattern="PERSON")
attr:answer_type == answer_type(prediction(model="bidaf"))
exact_match(model="bidaf") == 0
is_correct_sent(prediction(model="bidaf")) == 0
overlap(question, sentence(prediction(model="bidaf")) > overlap(question, sentence(groundtruths))
Prev page
Next page
Displaying #0-4 samples.
GROU...
all_inst...
is_entity
has_dis...
correct...
is_distr...
REWR...
STRIN...
keep_c...
remove...
resolve...


E
oups
Proportion
Show filtered distribution
Export the Groups
Compare models
all_instances 10570
is_entity length (ENT (groundtru 4240
has_distractor length (ENT (groundtru 3495
correct_type length (attr:ent_g) > 2988
is_distracted length (attr:ent_g) > 192
F
d Re-write Rules
Proportion
Show filtered distribution
Add a rule
Save rules
STRING (prediction (model="ANCHOR")) ≠ 202
keep_correct_sentence 92
remove_clues 17
resolve_coref 5

Building blocks

Quantitative grouping

Inspect similar instances,
semantically & syntactically

Counterfactual perturbation

Isolate important components
targeted minimal rewrites

Errudite

Precise & reproducible hypotheses
+
Scale up to the entire dev set
+
Cover **errors & correct** instances
+
Test via counterfactual analysis

Scenario: distractor hypothesis

Common error hypothesis in question answering, on **BiDAF** (Seo et al., 2016), with **SQuAD** (10570 instances; Rajpurkar et al., 2016)

Common belief: BiDAF…

- ✓ Matches entity types
  Knows to find a *PERSON*
- ✗ Finds the exact answer spans
  Distracted by other *PERSON* spans

Who created the 2005 theme for Doctor Who?

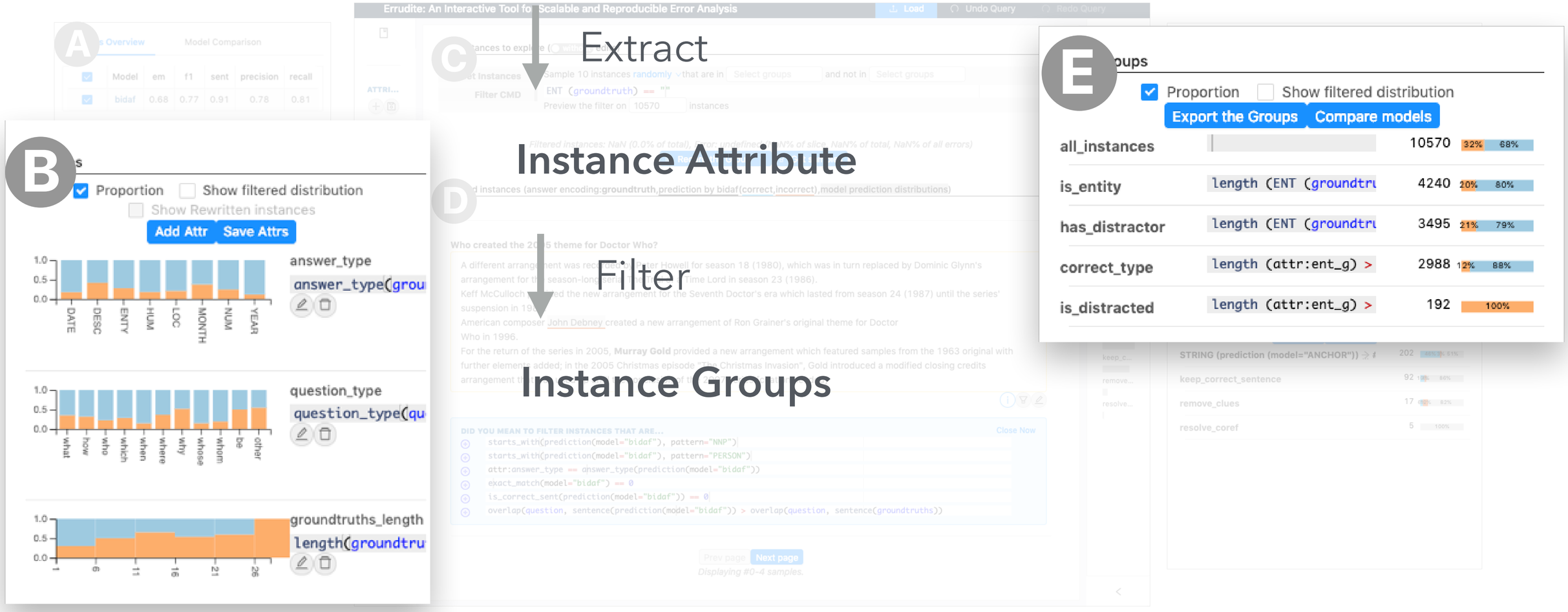
…**John Debney** created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996. For the return of the series in 2005, **Murray Gold** provided a new arrangement… sampled from the 1963 original.

Seo, M., Kembhavi, A., Farhadi, A., & Hajishirzi, H. (2016). Bidirectional attention flow for machine comprehension. ICLR 2017
Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016). Squad: 100,000+ questions for machine comprehension of text. EMNLP 2016

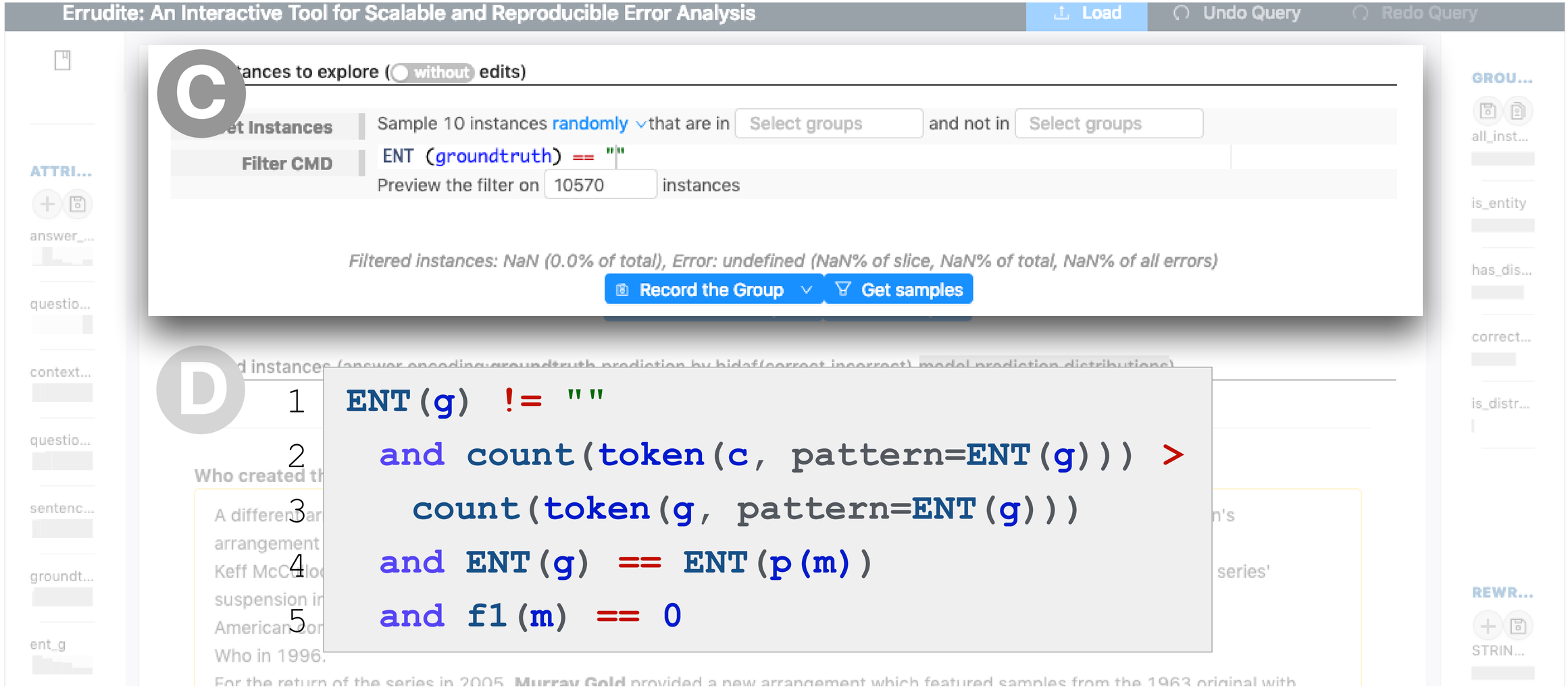
Precise DSL (Domain Specific Language)

Attribute Extractor + Target + Operators

```
length(q) > 20
```



Build distractor groups **with DSL**



Build distractor groups **with DSL**

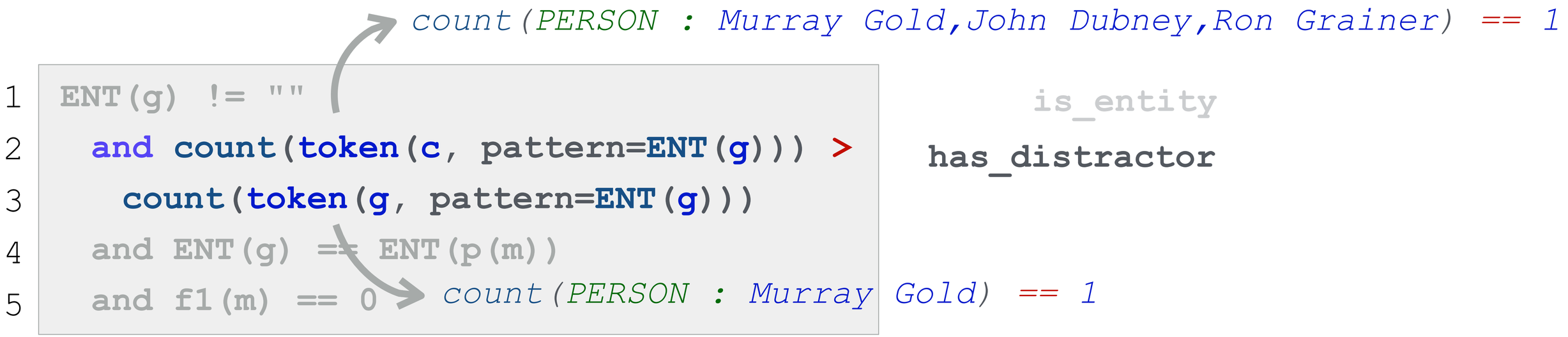
ENT(Murray Gold) == PERSON

```
1  ENT(g) != ""
2    and count(token(c, pattern=ENT(g))) >
3      count(token(g, pattern=ENT(g)))
4    and ENT(g) == ENT(p(m))
5    and f1(m) == 0
```

is_entity

*"The **g**roundtruth is an **ENT**ity."*

Build distractor groups **with DSL**



*"There are more tokens matching the ground truth entity type (**ENT(g)**) in the whole **c**ontext than in the **g**roundtruth."*

Build distractor groups **with DSL**

```
1  ENT(g) != ""
2    and count(token(c, pattern=ENT(g))) >
3      count(token(g, pattern=ENT(g)))
4    and ENT(g) == ENT(p(m))
5    and f1(m) == 0
```

```
is_entity
has_distractor
correct_type
```

ENT(John Debney) == PERSON

*"The **m**odel **p**rediction **ENT**ity type matches the **g**roundtruth **ENT**ity type."*

Build distractor groups **with DSL**

```
1  ENT(g) != ""
2    and count(token(c, pattern=ENT(g))) >
3      count(token(g, pattern=ENT(g)))
4    and ENT(g) == ENT(p(m))
5    and f1(m) == 0
```

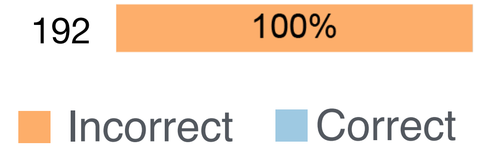
```
is_entity
has_distractor
correct_type
is_distracted
```

*"The **m**odel prediction is incorrect."*

Build distractor groups **with DSL**

```
1  ENT(g) != ""
2    and count(token(c, pattern=ENT(g))) >
3      count(token(g, pattern=ENT(g)))
4    and ENT(g) == ENT(p(m))
5    and f1(m) == 0
```

is_entity
has_distractor
correct_type
is_distracted

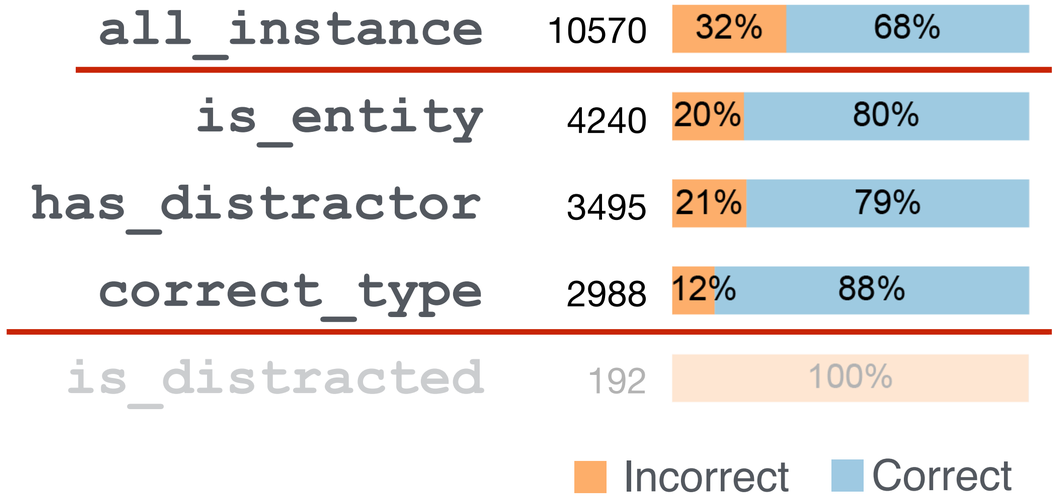


5.7% of all BiDAF errors:

The distractor hypothesis seems correct!

Build distractor groups **with DSL**

```
1  ENT(g) != ""
2    and count(token(c, pattern=ENT(g))) >
3      count(token(g, pattern=ENT(g)))
4    and ENT(g) == ENT(p(m))
5    and f1(m) == 0
```



88% EM > **68%** EM:

BiDAF performs better when have distractors & entity type is matched, than overall.

Reject / revise the hypothesis!

Building blocks

Quantitative grouping

Inspect similar instances, semantically & syntactically

Counterfactual perturbation

Isolate important components targeted minimal rewrites

Errudite

Precise & reproducible hypotheses

+

Scale up to the entire dev set

+

Cover **errors & correct** instances

+

Test via counterfactual analysis

Scenario: distractor hypothesis

`is_distracted` 192 100%

HAS distractor prediction

≠

IS WRONG due to distractor prediction

Who created the 2005 theme for Doctor Who?

…**John Debney** created a new arrangement of Ron Grainer's original theme for Doctor Who in 1996. For the return of the series in 2005, **Murray Gold** provided a new arrangement… sampled from the 1963 original.

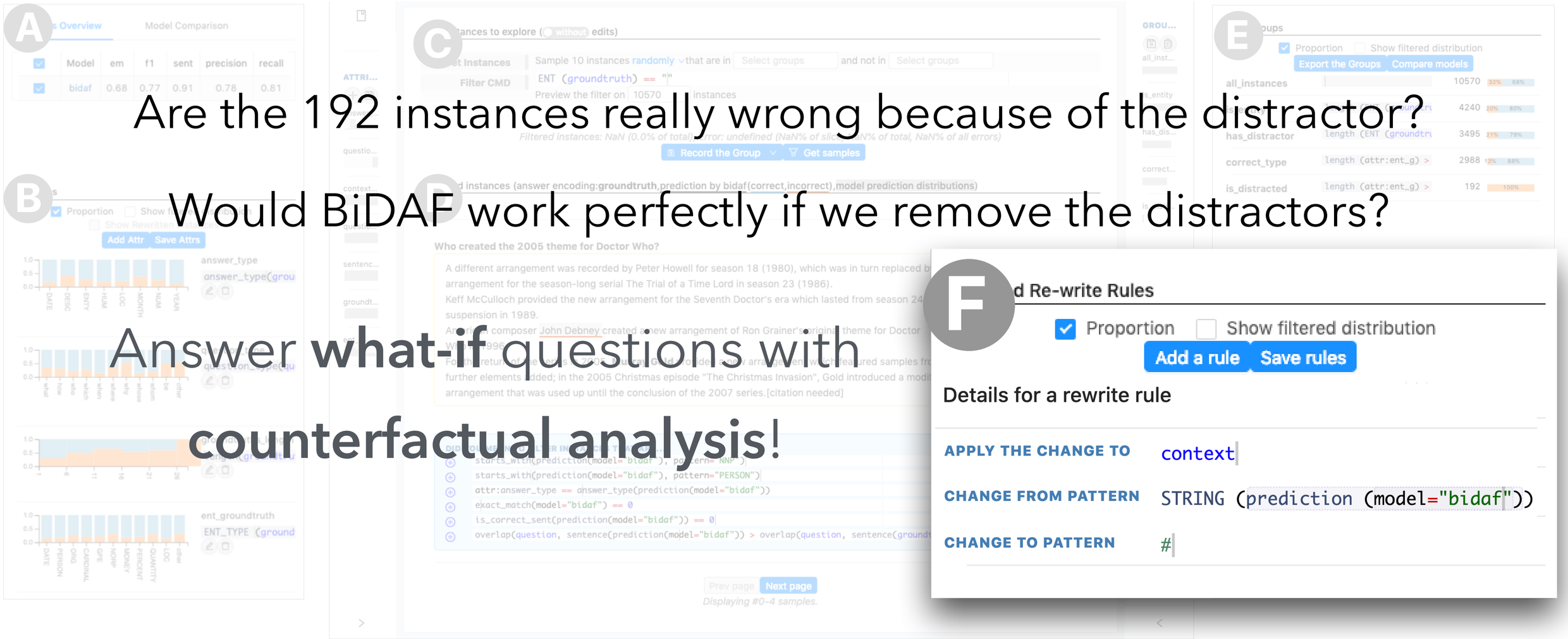
Distractor entity?

Multi-sentence reasoning?

Are the 192 instances really wrong because of the distractor?

Would BiDAF work perfectly if we remove the distractors?

Answer **what-if** questions with **counterfactual analysis**!



Counterfactual Analysis with Rewrite Rules

`rewrite(`**`target , from → to`**`)`

Re-write the **target** part of an instance

by replacing **from**

with **to**

Counterfactual Analysis with Rewrite Rules

Would BiDAF work perfectly if we **remove** the distractors?

`rewrite(`**`target , from → to`**`)`

Re-write the **target** part of an instance

by replacing **from**

with **to**

Counterfactual Analysis with Rewrite Rules

Would BiDAF work perfectly if we **remove** the distractors?

`rewrite( c , from → to )`

Re-write the **context** part of an instance

by replacing **from**

with **to**

Counterfactual Analysis with Rewrite Rules

Would BiDAF work perfectly if we **remove** the distractors?

`rewrite( c , string(p(m)) → to )`

Re-write the **context** part of an instance

by replacing the **model predicted distractor string**

with **to**

Counterfactual Analysis with Rewrite Rules

Would BiDAF work perfectly if we **remove** the distractors?

```
rewrite(   c    , string(p(m))   →   "#" )
```

Re-write the **context** part of an instance

by replacing the **model predicted distractor string**

with **a placeholder token "#"**

Counterfactual Analysis with Rewrite Rules

Would BiDAF work perfectly if we **remove** the distractors?

```
rewrite(   c    , string(p(m))   →   "#" )
```

Incorrect
Incorrect

Q: Who created the 2005 theme for Doctor Who?

C: …~~John Dobney~~ # created a new arrangement of Ron Grainer's …
Murray Gold provided a new arrangement…

Counterfactual Analysis with Rewrite Rules

Would BiDAF work perfectly if we **remove** the distractors?

```
rewrite(   c    , string(p(m))   →   "#" )
```

Incorrect
Incorrect

Another distractor is still confusing the model!

Counterfactual Analysis with Rewrite Rules

```
rewrite(    c     , string(p(m))   →   "#" )
```

p(m) for the 192 rewritten **is_distracted** instances are…

Incorrect
Incorrect

Another distractor is still confusing the model!

Counterfactual Analysis with Rewrite Rules

```
rewrite(   c    , string(p(m))   →   "#" )
```

p(m) for the 192 rewritten **is_distracted** instances are…

Incorrect → Incorrect	29%	✓ Another distractor is still confusing the model!
Incorrect → Correct	48%	✓ The distractor was fooling the model!
Unchanged	23%	✗ Other factors are at play!

age of 18, ~~10.5%~~ # from 18 to 24…

Building blocks

Quantitative grouping

Inspect similar instances, semantically & syntactically

Counterfactual perturbation

Isolate important components targeted minimal rewrites

Errudite

Precise & reproducible hypotheses

+

Scale up to the entire dev set

+

Cover **errors & correct** instances

+

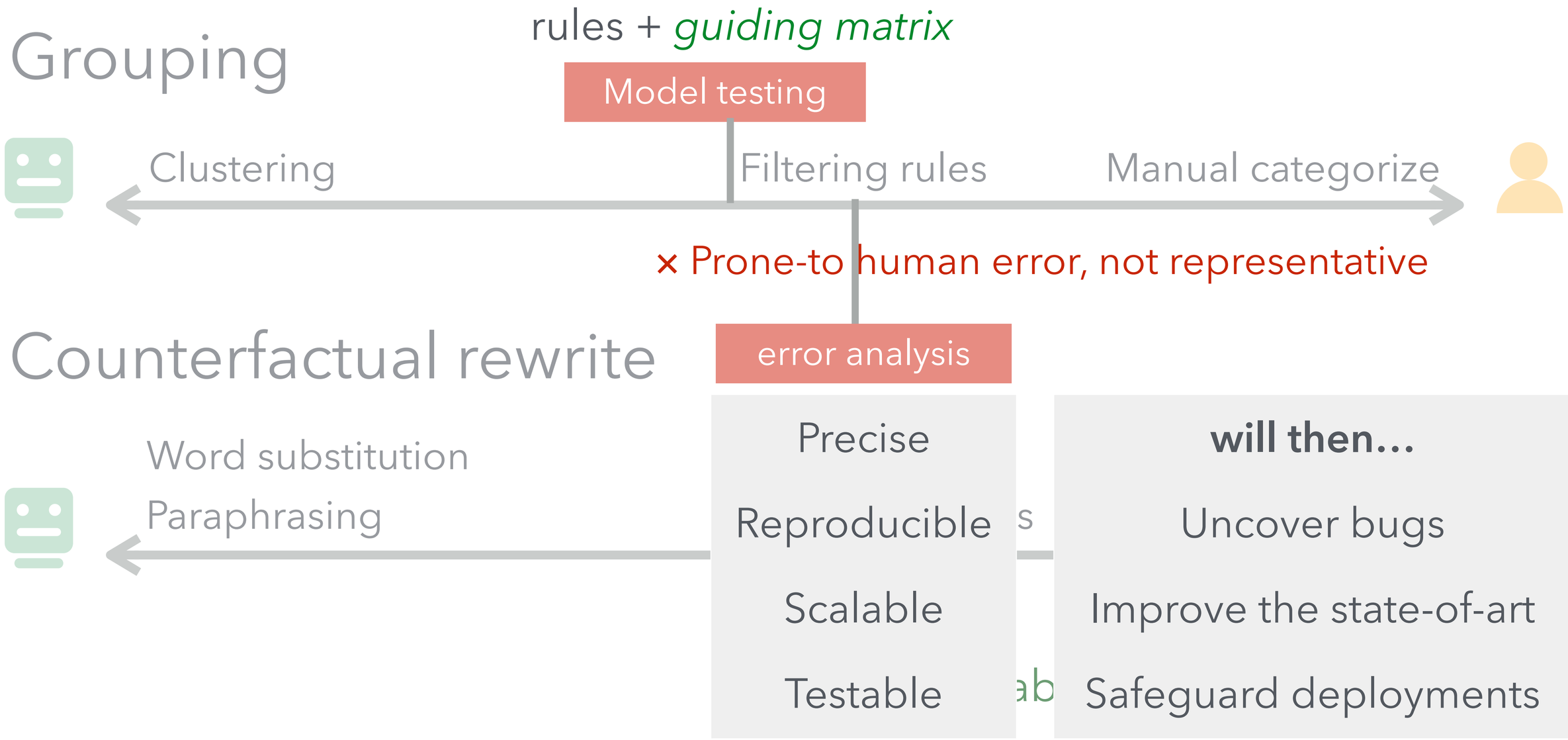
Test via counterfactual analysis

Grouping

Counterfactual rewrite

Errudite

√ scalable, interpretable


Grouping
rules + guiding matrix
Model testing
Clustering
Filtering rules
Manual categorize
× Prone-to human error, not representative
Counterfactual rewrite
error analysis
Word substitution
Paraphrasing
Precise
Reproducible
Scalable
Testable
will then…
Uncover bugs
Improve the state-of-art
Safeguard deployments

ACL 2020 Best Paper Award

Beyond Accuracy: Behavioral Testing of NLP Models with **Checklist**

Marco Tulio Ribeiro
Microsoft Research

Tongshuang (Sherry) Wu @tongshuangwu
University of Washington

Jeffrey Heer @jeffrey_heer
Daniel S. Weld @dsweld
University of Washington

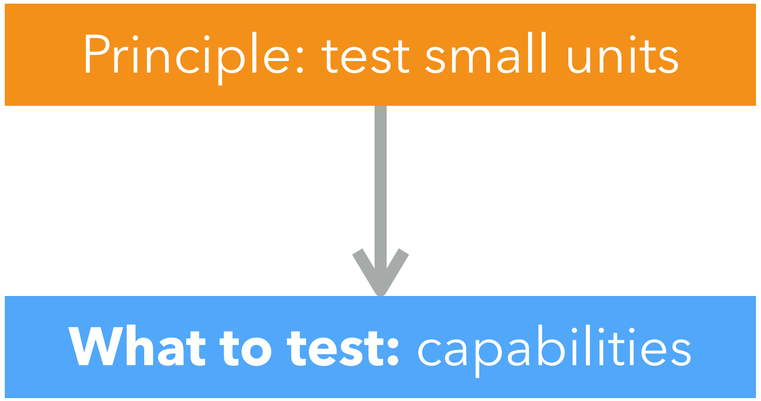



CheckList — Framework + Tooling

Applying the principles for Software Engineering testing to NLP

Software engineering → NLP

Capabilities	Descriptions
Vocab/POS	important words or word types for the task.
Named entities	appropriately understanding named entities.
Nagation	understand the negation words.
Taxonomy	synonyms, antonyms, etc.
Robustness	to typos, irrelevant changes, etc.
Coreference	resolve ambiguous pronouns, etc.
Fairness	not biasing towards certain gender/race groups.
Semantic Role Labeling	understanding roles such as agent, object, etc.
Logic	handle symmetry, consistency, and conjunctions.
Temporal	understand order of events.



Why do we have the universal list?

Models' required capabilities are task-independent.

Models' expected behaviors w.r.t capabilities are task-dependent.

This is not an exhaustive list!

Software engineering → NLP

Capabilities
Vocab/POS
Named entities
Nagation
…

Behavioral testing: decouple tests from implementation

↓

Decouple tests from training

Meets users' needs

Works with black box models

Software engineering → NLP

Capabilities			
Vocab/POS			
Named entities			
Nagation			
…			

Behavioral testing: decouple tests from implementation

↓

Decouple tests from training

How to test:
Test behaviors with different test types!

Illustrating task: **sentiment analysis** with **Google Cloud's Natural Language**

Software engineering → NLP

Capabilities	MFT		
Vocab/POS			
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Software engineering → NLP

Capabilities	MFT		
Vocab/POS			
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

} *A group of n=500 test cases*

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15%	← *1 test, with failure rate*	
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

} *A group of n=500 test cases*

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15%		
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

Expectation: Exact labels

This is a commercial flight. (neutral)

I flew to Indiana yesterday. (neutral)

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%	← *multiple tests per cell*	
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

Expectation: Exact labels

This is a commercial flight. (neutral)

I flew to Indiana yesterday. (neutral)

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

The cabin crew was not great. (negative)

I can't say I enjoyed the food. (negative)

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

The cabin crew was not great. (negative)

I can't say I enjoyed the food. (negative)

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

~~Start from scratch~~ → Perturb existing ones

~~Expect exact label~~ → Expect predictions to (not) change

Software engineering → NLP

Capabilities	MFT	INV	
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

Software engineering → NLP

Capabilities	MFT	INV	
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

No need to specify the exact prediction!

Expectation: Same prediction after the change.

@AmericanAir thank you we got on a different flight to ~~Chicago~~ Dallas.

@VirginAmerica I can't lose my luggage, moving to ~~Brazil~~ Turkey soon.

Software engineering → NLP

Capabilities	MFT	INV	
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

No need to specify the exact prediction!

Expectation: Same prediction after the change.

@AmericanAir thank you we got on a different flight to ~~Chicago~~ Dallas.

@VirginAmerica I can't lose my luggage, moving to ~~Brazil~~ Turkey soon.

Software engineering → NLP

Capabilities	MFT	INV	DIR
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

INVariance Tests

DIRectional Expectation Tests

Software engineering → NLP

Capabilities	MFT	INV	DIR
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

DIRectional Expectation Tests

Expectation: Sentiment monotonic decreasing (↓)

@AmericanAir service wasn't great. You are lame.

@JetBlue why won't YOU help them?! Ugh. I dread you.

expectation on probability!

Software engineering → NLP

Capabilities	MFT	INV	DIR
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		Add neg: 34.6%
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

DIRectional Expectation Tests

Expectation: Sentiment monotonic decreasing (↓)

@AmericanAir service wasn't great. You are lame.

@JetBlue why won't YOU help them?! Ugh. I dread you.

expectation on probability!

NLP testing in a nutshell: fill in the matrix

Tests are **grouped by** (capability, test type, expectation).

how?

what?

Capabilities	MFT	INV	DIR
Vocab/POS	✓	✘	✘
Named entities	✓	✓	✘
Nagation	✘	✓	✘
…			

Find a cell of (cap, test type)

Define (maybe ≥ 1) tests

test = test case + expectation

Run the model, get passes/fails

Form a test suite – reuse for other models!

CheckList — Framework + Tooling

Abstractions that ease the pain of the test generation, increase coverage.

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

CheckList as a tool

Templates

RoBERTa suggestions

Lexicons

Perturbation library

Expectation functions

Test inspecting/sharing

Visualization

More in our repo!

https://github.com/marcotcr/checklist

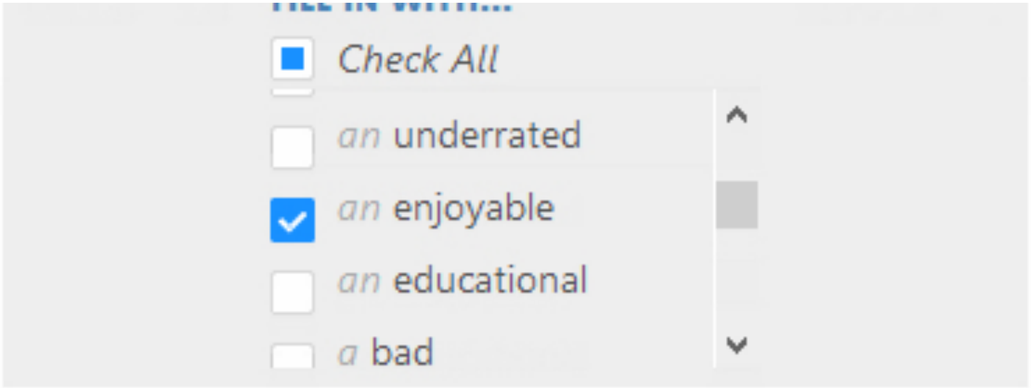
CheckList as a tool

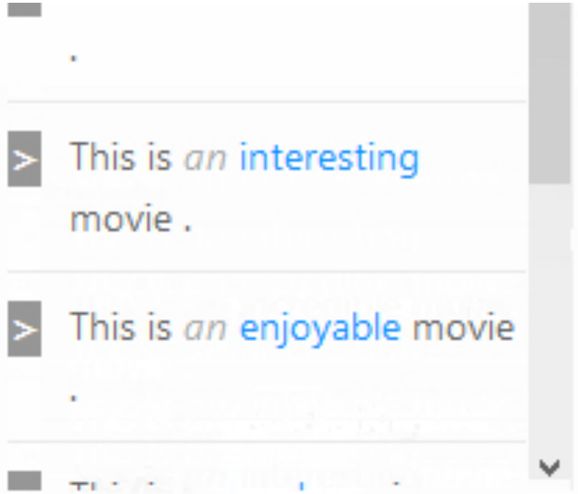
```
In [27]: editor.visual_suggest('This is {a:mask} movie.')
```


```
In [26]: editor.selected_suggestions
```

Wordnet

CheckList as a tool

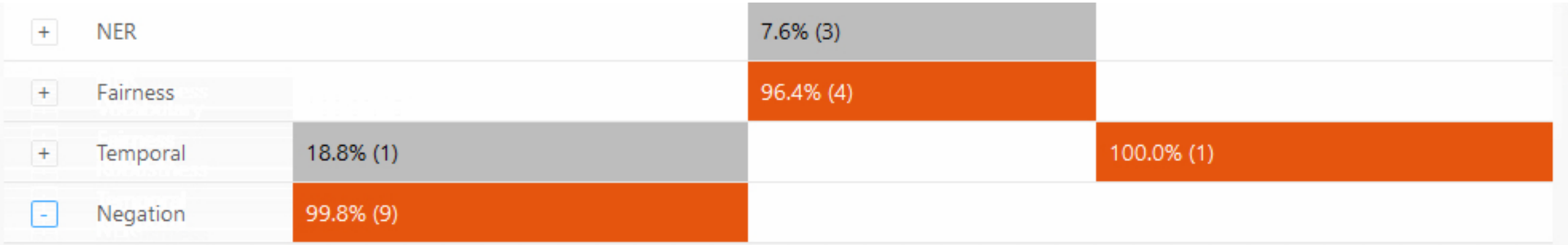





```
In [28]: editor.selected_suggestions

Out[28]: ['good',
          'amazing',
          'excellent',
          'interesting',
          'awesome',
          'incredible',
          'great',
          'enjoyable']
```

Capabilities	**M**inimum **F**unctionality **T**est *failure rate % (over N tests)*	**INV**ariance Test *failure rate % (over N tests)*	**DIR**ectional Expectation Test *failure rate % (over N tests)*
+ Vocabulary	100.0% (5)	10.2% (1)	0.8% (4)
+ Robustness		11.4% (5)	
+ NER		7.6% (3)	
+ Fairness		96.4% (4)	
+ Temporal	18.8% (1)		100.0% (1)
+ Negation	99.8% (9)		
+ SRL	100.0% (5)		



Minimum Functionality Test

	test name	failure rate
+	simple negations: negative	42 / 500 = 8.4%
+	simple negations: not negative	66 / 500 = 13.2%
+	simple negations: not neutral is still neutral	492 / 500 = 98.4%
+	simple negations: I thought x was positive, but it was not (should be negative)	11 / 500 = 2.2%
+	simple negations: I thought x was negative, but it was not (should be neutral or positive)	424 / 500 = 84.8%
+	simple negations: but it was not (neutral) should still be neutral	493 / 500 = 98.6%
+	Hard: Negation of positive with neutral stuff in the middle (should be negative)	370 / 500 = 74.0%
+	Hard: Negation of negative with neutral stuff in the middle	

+	simple negations: not negative	66 / 500 = 13.2%
+	simple negations: not neutral is still neutral	492 / 500 = 98.4%
+	simple negations: I thought x was positive, but it was not (should be negative)	11 / 500 = 2.2%
+	simple negations: I thought x was negative, but it was not (should be neutral or positive)	424 / 500 = 84.8%
+	simple negations: but it was not (neutral) should still be neutral	493 / 500 = 98.6%
-	Hard: Negation of positive with neutral stuff in the middle (should be negative)	370 / 500 = 74.0%

Test Summary

Test [MFT] on [NEGATION]

Hard: Negation of positive with neutral stuff in the middle (should be negative)

Result FAILURE RATE ON ALL CASES

370/500=74.0%

FILTER TEST CASES

Input free text and enter

Examples

Failed cases only

I would n't say , given that I am from Brazil , that this food was extraordinary .	Expect: 0 | Pred: 2 (1.00)
I would n't say , given it 's a Tuesday , that that is a beautiful aircraft .	Expect: 0 | Pred: 2 (1.00)
I would n't say , given that I am from Brazil , that the service is wonderful .	Expect: 0 | Pred: 2 (1.00)
I ca n't say , given all that I 've seen	Expect: 0 | Pred: 2 (1.00)

+	Hard: Negation of negative with neutral stuff in the middle	499 / 500 = 99.8%

+	simple negations: not negative	66 / 500 =	13.2%
+	simple negations: not neutral is still neutral	492 / 500 =	98.4%
+	simple negations: I thought x was positive, but it was not (should be negative)	11 / 500 =	2.2%
+	simple negations: I thought x was negative, but it was not (should be neutral or positive)	424 / 500 =	84.8%
+	simple negations: but it was not (neutral) should still be neutral	493 / 500 =	98.6%
-	Hard: Negation of positive with neutral stuff in the middle (should be negative)	370 / 500 =	74.0%

Test Summary

Test [MFT] on [NEGATION]

Hard: Negation of positive with neutral stuff in the middle (should be negative)

Result FAILURE RATE ON ALL CASES

370/500=74.0%

FILTER TEST CASES

Input free text and enter

Examples

Failed cases only

>	I would n't say , given that I am from Brazil , that this food was extraordinary .	Expect: 0 | Pred: 2 (1.00)
>	I would n't say , given it 's a Tuesday , that that is a beautiful aircraft .	Expect: 0 | Pred: 2 (1.00)
>	I would n't say , given that I am from Brazil , that the service is wonderful .	Expect: 0 | Pred: 2 (1.00)
>	I ca n't say , given all that I 've seen	Expect: 0 | Pred: 2 (1.00)

+	Hard: Negation of negative with neutral stuff in the middle	499 / 500 =	99.8%

+	simple negations: not negative	66 / 500 = 13.2%
+	simple negations: not neutral is still neutral	492 / 500 = 98.4%
+	simple negations: I thought x was positive, but it was not (should be negative)	11 / 500 = 2.2%
+	simple negations: I thought x was negative, but it was not (should be neutral or positive)	424 / 500 = 84.8%
+	simple negations: but it was not (neutral) should still be neutral	493 / 500 = 98.6%
-	Hard: Negation of positive with neutral stuff in the middle (should be negative)	370 / 500 = 74.0%

Test Summary

Test [MFT] on [NEGATION]

Hard: Negation of positive with neutral stuff in the middle (should be negative)

Result FAILURE RATE ON ALL CASES

370/500=74.0%

FILTER TEST CASES

Input free text and enter

Examples

Failed cases only

- I would n't say , given that I am from Brazil , that this food was extraordinary . — Expect: 0 | Pred: 2 (1.00)
- I would n't say , given it 's a Tuesday , that that is a beautiful aircraft . — Expect: 0 | Pred: 2 (1.00)
- I would n't say , given that I am from Brazil , that the service is wonderful . — Expect: 0 | Pred: 2 (1.00)
- I ca n't say , given all that I 've seen — Expect: 0 | Pred: 2 (1.00)

+	Hard: Negation of negative with neutral stuff in the middle	499 / 500 = 99.8%

+	simple negations: not negative	66 / 500 = 13.2%
+	simple negations: not neutral is still neutral	492 / 500 = 98.4%
+	simple negations: I thought x was positive, but it was not (should be negative)	11 / 500 = 2.2%
+	simple negations: I thought x was negative, but it was not (should be neutral or positive)	424 / 500 = 84.8%
+	simple negations: but it was not (neutral) should still be neutral	493 / 500 = 98.6%
-	Hard: Negation of positive with neutral stuff in the middle (should be negative)	370 / 500 = 74.0%

Test Summary

Test [MFT] on [NEGATION]
Hard: Negation of positive with neutral stuff in the middle (should be negative)

Result

FAILURE RATE ON ALL CASES

370/500=74.0%

FILTER TEST CASES

Input free text and enter

Examples

Failed cases only

- I would n't say , given that I am from Brazil , that this food was extraordinary . — Expect: 0 | Pred: 2 (1.00)
- I would n't say , given it 's a Tuesday , that that is a beautiful aircraft . — Expect: 0 | Pred: 2 (1.00)
- I would n't say , given that I am from Brazil , that the service is wonderful . — Expect: 0 | Pred: 2 (1.00)
- I ca n't say , given all that I 've seen — Expect: 0 | Pred: 2 (1.00)

+	Hard: Negation of negative with neutral stuff in the middle	499 / 500 = 99.8%

This is too simple, you won't find any bugs.

Case Study & User Study

Let's test some SOTA models (that some people consider solved)!
sentiment analysis, QQP, QA

Sentiment analysis

Task Twitter sentiment analysis

@AmericanAir thank you for a delightful flight to Chicago! (positive)

Claimed to be a use case by all commercial models!

Models

Commercial models

Microsoft's Text Analytics

Google Cloud's Natural Language

Amazon's Comprehend

Research models

BERT (trained on SST-2)

RoBERTa (trained on SST-2)

.https://azure.microsoft.com/en-us/services/cognitive-services/text-analytics/

https://cloud.google.com/natural-language

https://aws.amazon.com/cn/comprehend/

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013, October). Recursive deep models for semantic compositionality over a sentiment treebank. In Proceedings of the 2013 conference on empirical methods in natural language processing (pp. 1631-1642).

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS			
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
…			

Replace neutral words with BERT

Inputs (n=500) & expectations

~~the~~ our nightmare continues **(INV)**

@Virgin should I be concerned ~~that~~ when I'm about to fly… **(INV)**

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✘	
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
…			

Replace neutral words with BERT

Inputs (n=500) & expectations

~~the~~ our nightmare continues **(INV)**

@Virgin should I be concerned ~~that~~ when I'm about to fly… **(INV)**

				**RoBERTa**
9.4	16.2	12.4	10.2	10.2

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✘	
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
…			

Add negative phrases

Inputs (n=500) & expectations

@SouthwestAir ok, gotcha! I abhor you (↓)

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✘	✘
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
…			

Add negative phrases

Inputs (n=500) & expectations

@SouthwestAir ok, gotcha! I abhor you (↓)

0.8	34.6	5.0	0.0	13.2

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✘	✘
Taxonomy			
Robustness			
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
…			

Add random url or @

Inputs (n=500) & expectations

@JetBlue that selfie was extreme. @pi9QDK **(INV)**

				**RoBERTa**
9.6	13.4	24.8	11.4	7.4

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✗	✗
Taxonomy			
Robustness		✗	
NER			
Fairness			
Temporal			
Nagation			
Coreference			
SRL			
Logic			
…			

Temporal change

Inputs (n=500) & expectations

I used to hate this airline, although now I like it **(Pos)**

				**RoBERTa**
41.0	36.6	42.2	18.8	11.0

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✗	✗
Taxonomy			
Robustness		✗	
NER			
Fairness			
Temporal	✗		
Nagation			
Coreference			
SRL			
Logic			
…			

Negated negation

Inputs (n=500) & expectations

It wasn't a lousy customer service **(Pos or Neutral)**

				RoBERTa
18.8	54.2	29.4	13.2	10.2

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✘	✘
Taxonomy			
Robustness		✘	
NER			
Fairness			
Temporal	✘		
Nagation	✘		
Coreference			
SRL			
Logic			
…			

Q&A form

Inputs (n=500) & expectations

Do I think this company is bad? No **(Pos or Neutral)**

				RoBERTa
96.8	90.8	81.6	55.4	54.8

Sentiment analysis

Capabilities	MFT	INV	DIR
Vocab/POS		✘	✘
Taxonomy			
Robustness		✘	
NER			
Fairness			
Temporal	✘		
Nagation	✘		
Coreference			
SRL	✘		
Logic			
…			

Q&A form

Inputs (n=500) & expectations

Do I think this company is bad? No **(Pos or Neutral)**

				RoBERTa
96.8	90.8	81.6	55.4	54.8

Case study: Microsoft Sentiment Analysis

Microsoft

Model already stress tested, continue to improve

Public benchmarks

In-house benchmarks (e.g. negation)

User complaint benchmarks

CheckList: 5 hour session

Find many new bugs

Test new capabilities

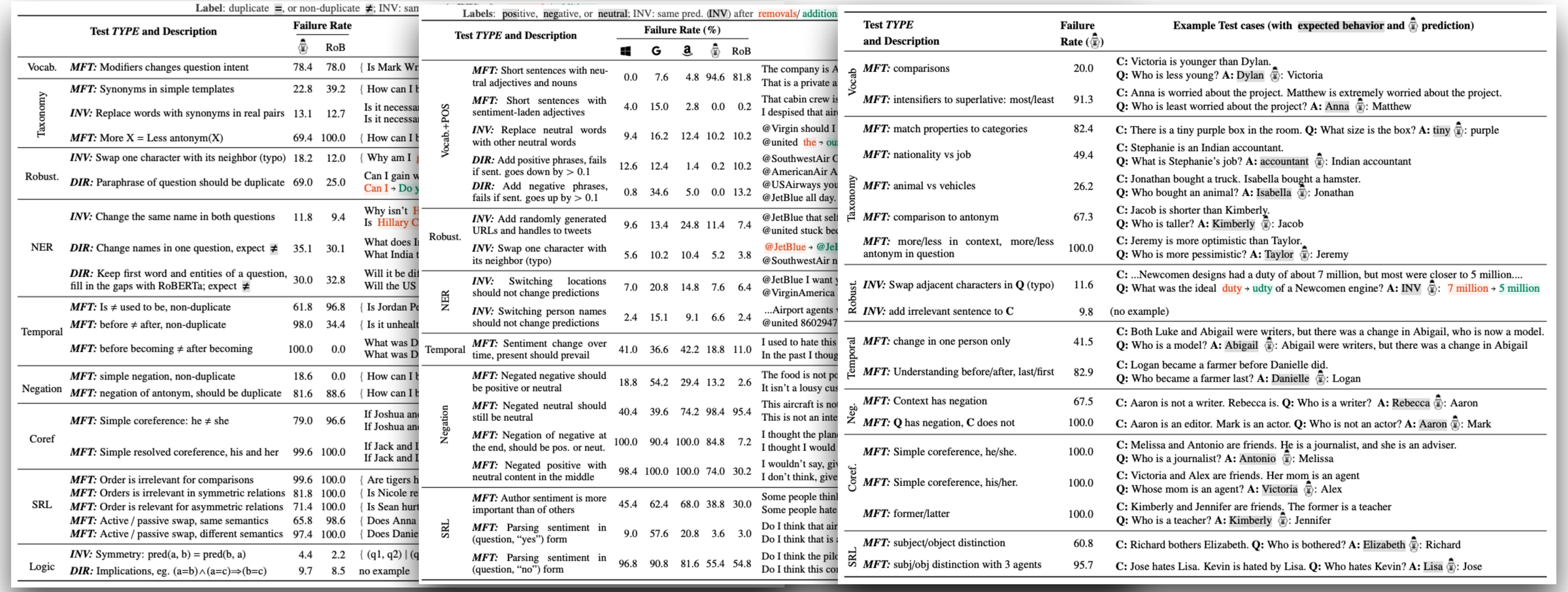
Test old capabilities better

Same process, different tasks & models

Sentiment analysis

Question Pair Detection

Question Answering



Label: duplicate (=), or non-duplicate (≠); INV: san

	Test *TYPE* and Description	Failure Rate		
			RoB	
Vocab.	***MFT:*** Modifiers changes question intent	78.4	78.0	{ Is Mark Wr
Taxonomy	***MFT:*** Synonyms in simple templates	22.8	39.2	{ How can I b
	INV: Replace words with synonyms in real pairs	13.1	12.7	Is it necessa Is it necessa
	MFT: More X = Less antonym(X)	69.4	100.0	{ How can I b
Robust.	***INV:*** Swap one character with its neighbor (typo)	18.2	12.0	{ Why am I
	DIR: Paraphrase of question should be duplicate	69.0	25.0	Can I gain w Can I → Do y
NER	***INV:*** Change the same name in both questions	11.8	9.4	Why isn't H Is Hillary C
	DIR: Change names in one question, expect ≠	35.1	30.1	What does I What India t
	DIR: Keep first word and entities of a question, fill in the gaps with RoBERTa; expect ≠	30.0	32.8	Will it be di Will the US
Temporal	***MFT:*** Is ≠ used to be, non-duplicate	61.8	96.8	{ Is Jordan Pe
	MFT: before ≠ after, non-duplicate	98.0	34.4	{ Is it unhealt
	MFT: before becoming ≠ after becoming	100.0	0.0	What was D What was D
Negation	***MFT:*** simple negation, non-duplicate	18.6	0.0	{ How can I b
	MFT: negation of antonym, should be duplicate	81.6	88.6	{ How can I b
Coref	***MFT:*** Simple coreference: he ≠ she	79.0	96.6	If Joshua an If Joshua an
	MFT: Simple resolved coreference, his and her	99.6	100.0	If Jack and I If Jack and I
SRL	***MFT:*** Order is irrelevant for comparisons	99.6	100.0	{ Are tigers h
	MFT: Orders is irrelevant in symmetric relations	81.8	100.0	{ Is Nicole re
	MFT: Order is relevant for asymmetric relations	71.4	100.0	{ Is Sean hurt
	MFT: Active / passive swap, same semantics	65.8	98.6	{ Does Anna
	MFT: Active / passive swap, different semantics	97.4	100.0	{ Does Danie
Logic	***INV:*** Symmetry: pred(a, b) = pred(b, a)	4.4	2.2	{ (q1, q2) | (q
	DIR: Implications, eg. (a=b)∧(a=c)⇒(b=c)	9.7	8.5	no example

Labels: positive, negative, or neutral; INV: same pred. (INV) after removals/ addition

	Test *TYPE* and Description	Failure Rate (%)					
			G			RoB	
Vocab.+POS	***MFT:*** Short sentences with neutral adjectives and nouns	0.0	7.6	4.8	94.6	81.8	The company is A That is a private a
	MFT: Short sentences with sentiment-laden adjectives	4.0	15.0	2.8	0.0	0.2	That cabin crew is I despised that air
	INV: Replace neutral words with other neutral words	9.4	16.2	12.4	10.2	10.2	@Virgin should I @united the → ou
	DIR: Add positive phrases, fails if sent. goes down by > 0.1	12.6	12.4	1.4	0.2	10.2	@SouthwestAir C @AmericanAir A
	DIR: Add negative phrases, fails if sent. goes up by > 0.1	0.8	34.6	5.0	0.0	13.2	@USAirways you @JetBlue all day.
Robust.	***INV:*** Add randomly generated URLs and handles to tweets	9.6	13.4	24.8	11.4	7.4	@JetBlue that sel @united stuck be
	INV: Swap one character with its neighbor (typo)	5.6	10.2	10.4	5.2	3.8	@JetBlue → @JeB @SouthwestAir n
NER	***INV:*** Switching locations should not change predictions	7.0	20.8	14.8	7.6	6.4	@JetBlue I want y @VirginAmerica
	INV: Switching person names should not change predictions	2.4	15.1	9.1	6.6	2.4	...Airport agents @united 8602947
Temporal	***MFT:*** Sentiment change over time, present should prevail	41.0	36.6	42.2	18.8	11.0	I used to hate this In the past I thoug
Negation	***MFT:*** Negated negative should be positive or neutral	18.8	54.2	29.4	13.2	2.6	The food is not pc It isn't a lousy cus
	MFT: Negated neutral should still be neutral	40.4	39.6	74.2	98.4	95.4	This aircraft is no This is not an inte
	MFT: Negation of negative at the end, should be pos. or neut.	100.0	90.4	100.0	84.8	7.2	I thought the plan I thought I would
	MFT: Negated positive with neutral content in the middle	98.4	100.0	100.0	74.0	30.2	I wouldn't say, giv I don't think, give
SRL	***MFT:*** Author sentiment is more important than of others	45.4	62.4	68.0	38.8	30.0	Some people thin Some people hate
	MFT: Parsing sentiment in (question, "yes") form	9.0	57.6	20.8	3.6	3.0	Do I think that air Do I think that is
	MFT: Parsing sentiment in (question, "no") form	96.8	90.8	81.6	55.4	54.8	Do I think the pilo Do I think this co

	Test *TYPE* and Description	Failure Rate	Example Test cases (with expected behavior and prediction)
Vocab	***MFT:*** comparisons	20.0	**C:** Victoria is younger than Dylan. **Q:** Who is less young? **A:** Dylan: Victoria
	MFT: intensifiers to superlative: most/least	91.3	**C:** Anna is worried about the project. Matthew is extremely worried about the project. **Q:** Who is least worried about the project? **A:** Anna: Matthew
Taxonomy	***MFT:*** match properties to categories	82.4	**C:** There is a tiny purple box in the room. **Q:** What size is the box? **A:** tiny: purple
	MFT: nationality vs job	49.4	**C:** Stephanie is an Indian accountant. **Q:** What is Stephanie's job? **A:** accountant: Indian accountant
	MFT: animal vs vehicles	26.2	**C:** Jonathan bought a truck. Isabella bought a hamster. **Q:** Who bought an animal? **A:** Isabella: Jonathan
	MFT: comparison to antonym	67.3	**C:** Jacob is shorter than Kimberly. **Q:** Who is taller? **A:** Kimberly: Jacob
	MFT: more/less in context, more/less antonym in question	100.0	**C:** Jeremy is more optimistic than Taylor. **Q:** Who is more pessimistic? **A:** Taylor: Jeremy
Robust.	***INV:*** Swap adjacent characters in **Q** (typo)	11.6	**C:** ...Newcomen designs had a duty of about 7 million, but most were closer to 5 million.... **Q:** What was the ideal duty → udty of a Newcomen engine? **A:** INV: 7 million → 5 million
	INV: add irrelevant sentence to **C**	9.8	(no example)
Temporal	***MFT:*** change in one person only	41.5	**C:** Both Luke and Abigail were writers, but there was a change in Abigail, who is now a model. **Q:** Who is a model? **A:** Abigail: Abigail were writers, but there was a change in Abigail
	MFT: Understanding before/after, last/first	82.9	**C:** Logan became a farmer before Danielle did. **Q:** Who became a farmer last? **A:** Danielle: Logan
Neg.	***MFT:*** Context has negation	67.5	**C:** Aaron is not a writer. Rebecca is. **Q:** Who is a writer? **A:** Rebecca: Aaron
	MFT: **Q** has negation, **C** does not	100.0	**C:** Aaron is an editor. Mark is an actor. **Q:** Who is not an actor? **A:** Aaron: Mark
Coref.	***MFT:*** Simple coreference, he/she.	100.0	**C:** Melissa and Antonio are friends. He is a journalist, and she is an adviser. **Q:** Who is a journalist? **A:** Antonio: Melissa
	MFT: Simple coreference, his/her.	100.0	**C:** Victoria and Alex are friends. Her mom is an agent **Q:** Whose mom is an agent? **A:** Victoria: Alex
	MFT: former/latter	100.0	**C:** Kimberly and Jennifer are friends. The former is a teacher **Q:** Who is a teacher? **A:** Kimberly: Jennifer
SRL	***MFT:*** subject/object distinction	60.8	**C:** Richard bothers Elizabeth. **Q:** Who is bothered? **A:** Elizabeth: Richard
	MFT: subj/obj distinction with 3 agents	95.7	**C:** Jose hates Lisa. Kevin is hated by Lisa. **Q:** Who hates Kevin? **A:** Lisa: Jose

See our paper!

Building blocks

Quantitative grouping

Inspect similar instances, semantically & syntactically

Counterfactual perturbation

Isolate important components targeted minimal rewrites

Checklist

What to test

Capabilities, shared across tasks

+ How to test

Simple examples (MFTs), perturbations (INVs, DIRs)

+ Tooling

BERT fill-ins, visualizations, lexicons, . . .

How to fix bugs found in CheckList?

Perturbations as feedback to model training, dataset augmentation, etc.

Data manage. | Model verification

augment | perform - explanation

Submitted to ACL 2021

Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving models

Tongshuang (Sherry) Wu @tongshuangwu
University of Washington

Marco Tulio Ribeiro
Microsoft Research

Jeffrey Heer @jeffrey_heer
Daniel S. Weld @dsweld
University of Washington




Most work!

Word substitution

Paraphrasing

Errudite

CheckList

Snorkel

Challenge sets

Heuristic rules

Rewrite functions

Counterfactual data aug.

Contrast Set

Manual rewrite

× limited patterns

√ Scalable

√ Middle ground!

× Domain expertise

× Tedious

did → didn't

would never → would

√ Diverse

× Costly to scale

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. ACL 2020

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. Snorkel: Rapid training data creation with weak supervision, VLDB 2017.

Gardner, M., Artzi, Y., Basmov, V., Berant, J., Bogin, B., Chen, S., ... & Gupta, N. Evaluating models' local decision boundaries via contrast sets. EMNLP 2020

Kaushik, D., Hovy, E., & Lipton, Z. C. (2019). Learning the difference that makes a difference with counterfactually-augmented data. ICLR 2020.

Counterfactual rewriting as an NLG task

Prompt: I *did* like the movie.
Generate: I *didn't* like the movie

Language models can complete paragraphs → be finetuned for rewriting.

More **diverse** patterns from models like GPT-2

Design prompts to teach the model "**how** " and "**where**" to change

did → didn't, would never → would → *"add negation modifiers to aux"*

how *where*

“How to change”: Control codes

Control codes	Definition, color: delete → insert

“How to change”: Control codes

Constraints

concrete

Control codes	Definition, color: delete → insert
`negation`	A dog is not embraced by the woman.
`quantifier`	A dog is → Three dogs are embraced by the woman.
`shuffle`	*To move (or swap) key phrases or entities around the sentence.* A dog → woman is embraced by the woman → dog.

"How to change": Control codes

Constraints

concrete → loose

Control codes	Definition, color: delete → insert
`negation`	A dog is not embraced by the woman.
`quantifier`	A dog is → Three dogs are embraced by the woman.
`shuffle`	*To move (or swap) key phrases or entities around the sentence.* A dog → woman is embraced by the woman → dog.
`lexical`	*Changing just one word or noun chunks without breaking the POS tags.* A dog is embraced → attacked by the woman.
`resemantic`	*To replace short phrases or clauses without affecting the parsing tree.* A dog is embraced by the woman → wrapped in a blanket.

"How to change": Control codes

Constraints

concrete → loose

Group by

syntactic > semantic

Control codes	Definition, color: delete → insert
`negation`	A dog is not embraced by the woman.
`quantifier`	A dog is → Three dogs are embraced by the woman.
`shuffle`	*To move (or swap) key phrases or entities around the sentence.* A dog → woman is embraced by the woman → dog.
`lexical`	*Changing just one word or noun chunks without breaking the POS tags.* A dog is embraced → attacked by the woman.
`resemantic`	*To replace short phrases or clauses without affecting the parsing tree.* A dog is embraced by the woman → wrapped in a blanket.
`insert`	*To add constraints without affecting the parsing structure of other parts.* A dog is embraced by the little woman.
`delete`	*To remove constraints without affecting the parsing structure of other parts.* A dog is embraced by the woman.
`restructure`	*To alter the dependency tree structure, e.g. changing from passive to positive.* A dog is embraced by → hugging he woman.

“How to change”: Control codes

No training data?

Merge existing ones!

5 datasets →

191,415 sentence pairs

Control codes
negation
quantifier
shuffle
lexical
resemantic
insert
delete
restructure

PAWS, high lexical overlap, but non-paraphrasing sentences.
Can a bad → good person be good → bad?

WinoGrande, commonsense on lexical
The lions ate the zebras because they are predators → meaty.

HANS, challenge set for Natural Language Inference (NLI).
The banker near the judge saw the actor.

“Where to change”: Fill-in-the-blank (Donahue, ACL’20)

(It is great for kids., it is not great for children.)

It is not great for kids → children.

```
It is great for kids. <|perturb|> [negation]
  It is [BLANK] great for [BLANK]. [SEP] not [ANSWER] children [ANSWER]



<|endoftext|>
```

Donahue, C., Lee, M., & Liang, P. (2020). Enabling Language Models to Fill in the Blanks. ACL 2020

“Where to change”: Fill-in-the-blank (Donahue, ACL’20)

(It is great for kids., it is not great for children.)

It is not great for kids → children.

```
It is great for kids. <|perturb|> [negation]
  It is [BLANK] great for [BLANK]. [SEP] not [ANSWER] children [ANSWER]
  It [BLANK] great [BLANK]. [SEP]  is not [ANSWER] children [ANSWER]
  It is [BLANK]. [SEP] not great for children [ANSWER]
  [BLANK] [SEP] It is not great for children. [ANSWER]
<|endoftext|>
```

191,415 sentence pairs → 657,144 training prompts

Donahue, C., Lee, M., & Liang, P. (2020). Enabling Language Models to Fill in the Blanks. ACL 2020

Filter based on desired tasks

It is not great for kids → children.

It is great for anyone but kids.

It is great → disastrous for kids.

It is great → good for kids.

It is great → unnecessary for kids.

auto-augment

semantic dist

analyze “great”

It is great → good for kids.

It is great → disastrous for kids.

It is great → good for kids.

It is great → unnecessary for kids.

Counterfactual data augmentation

Does the model get more robust afterwards?

Similar to Kaushik et al.

Kaushik, D., Hovy, E., & Lipton, Z. C. (2019). Learning the difference that makes a difference with counterfactually-augmented data. ICLR 2020.

Counterfactual data aug: **crowd labeling**

Ranking: diversity!

prefer

It is great for anyone but kids.
It is great → disastrous for kids.

over

It is great → good for kids.
It is great → disastrous for kids.

Crowd labeling on MTurk

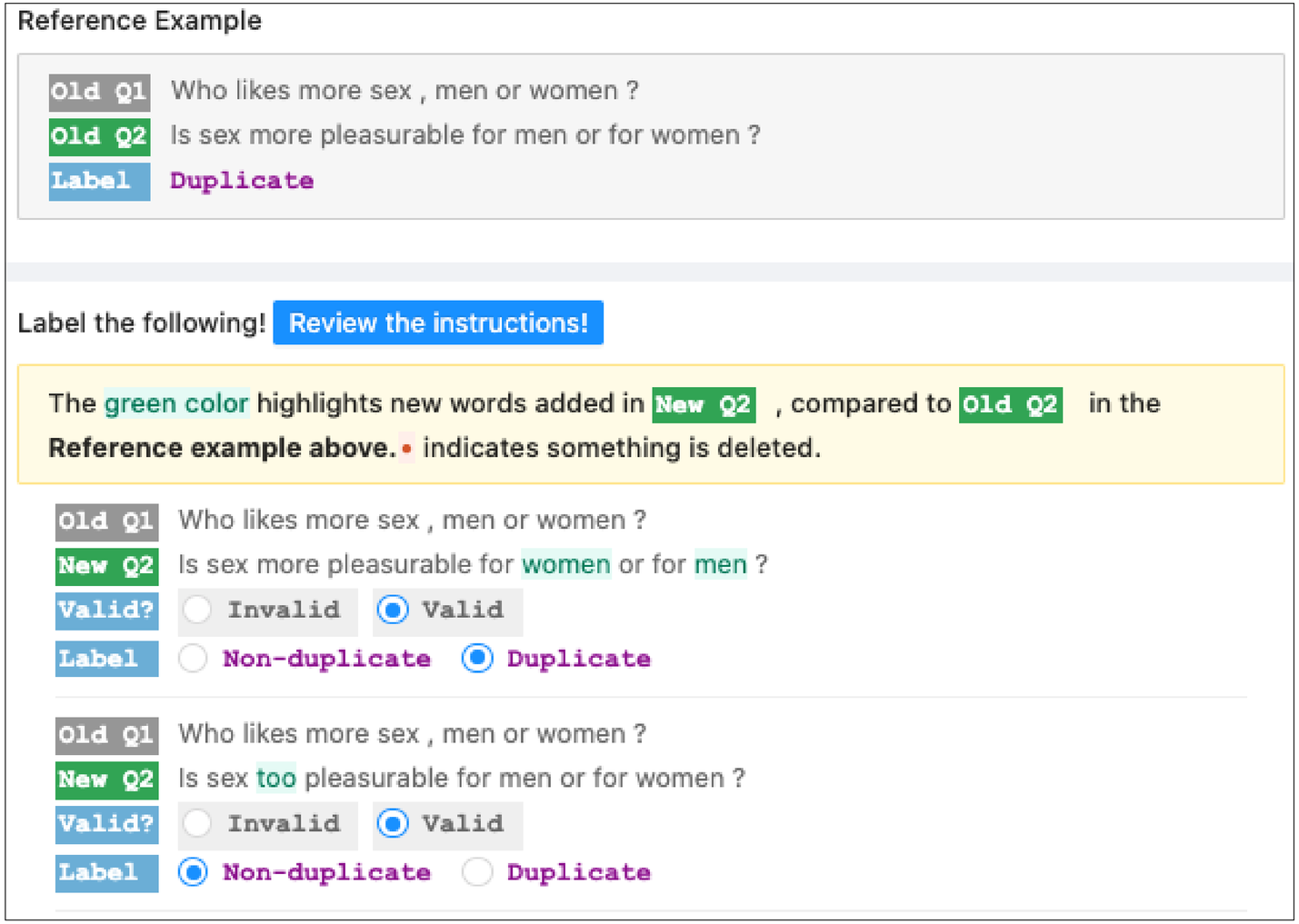
Validity? Class label?

Three tasks

Sentiment analysis

Natural Language Inference (NLI)

Duplicate question detection (QQP)



Counterfactual data aug: **Training results**

Sentiment analysis (on Stanford Sentiment Treebank)

I have not been this disappointed by → excited to see a movie in a long time. (negative → positive)

We just don't really care about this love story . (negative → positive)

`polyjuice`: 4k original + 2k counterfactuals (6k total)

v.s. **`baseline`** : 4k original + 2k extra, compensating original (6k total)

Model	SST-2	Senti140	SemEval	Amzbook	Yelp	IMDB	IMDB-Cont.	IMDB-CAD
`m-baseline`	92.9 ± 0.2	88.9 ± 0.3	84.8 ± 0.5	85.1 ± 0.4	90.0 ± 0.3	90.8 ± 0.5	92.2 ± 0.6	86.5 ± 0.2
`m-polyjuice`	92.7 ± 0.2	**90.7 ± 0.4**	**86.4 ± 0.1**	85.6 ± 0.8	90.1 ± 0.0	90.6 ± 0.3	**94.0 ± 0.3**	**89.7 ± 0.5**

human generated counterfactuals from prior papers

Improve on Twitter data (Senti140, SemEval), and contrast sets (IMDB-CDA, IMDB-Cont.)

Maintain accuracies on in domain, and other reviews data.

Counterfactual data aug: **Training results**

Natural Language Inference (on SNLI)

Premise: An airborne man on a surfboard.

Hypothesis: A man in mid air while his surfboard is beneath → lying on him. (entail → contradict)

`polyjuice`: 20k training + 1.5k perturbations, automatically gen. from polyjuice

v.s. **`CAD`** : 20k training + 1.5k perturbations, manually gen. from Kaushik et al.

v.s. **`baseline`** : 20k training + 1.5k extra, compensating original

Model	SNLI	MNLI-m	MNLI-mm	SNLI-CAD	break	DNC	stress	diagnostic
m-baseline	85.7 ± 0.4	86.1 ± 0.2	**86.6 ± 0.2**	72.8 ± 0.3	86.4 ± 1.5	54.5 ± 0.6	65.1 ± 0.6	56.0 ± 0.8
m-CAD	85.8 ± 0.6	**86.6 ± 0.1**	85.6 ± 0.3	**73.8 ± 0.2**	89.4 ± 2.9	55.8 ± 0.9	65.5 ± 0.5	56.4 ± 0.4
m-polyjuice	85.3 ± 0.3	86.0 ± 0.1	**86.4 ± 0.0**	**73.6 ± 0.2**	**89.1 ± 1.2**	**57.7 ± 0.3**	65.1 ± 0.2	**57.5 ± 0.5**

Similarly, improvements on multiple contrast/challenge sets, even better than CAD

Counterfactual data aug: **Training results**

Natural Language Inference (on SNLI)

Premise: An airborne man on a surfboard.

Hypothesis: A man in mid air while his surfboard is beneath → lying on him. (entail → contradict)

`polyjuice`: 20k training + 1.5k perturbations, automatically gen. from polyjuice

v.s. **`CAD`** : 20k training + 1.5k perturbations, manually gen. from Kaushik et al.

v.s. **`baseline`** : 20k training + 1.5k extra, compensating original

Model	SNLI	MNLI-m	MNLI-mm	SNLI-CAD	break	DNC	stress	diagnostic
m-baseline	85.7 ± 0.4	86.1 ± 0.2	**86.6 ± 0.2**	72.8 ± 0.3	86.4 ± 1.5	54.5 ± 0.6	65.1 ± 0.6	56.0 ± 0.8
m-CAD	85.8 ± 0.6	**86.6 ± 0.1**	85.6 ± 0.3	**73.8 ± 0.2**	89.4 ± 2.9	55.8 ± 0.9	65.5 ± 0.5	56.4 ± 0.4
m-polyjuice	85.3 ± 0.3	86.0 ± 0.1	**86.4 ± 0.0**	**73.6 ± 0.2**	**89.1 ± 1.2**	**57.7 ± 0.3**	65.1 ± 0.2	**57.5 ± 0.5**
m-polyjuice-rand	85.7 ± 0.4	86.1 ± 0.1	86.2 ± 0.1	73.4 ± 0.5	87.2 ± 0.6	54.7 ± 0.3	64.6 ± 0.6	56.9 ± 0.8

Similarly, improvements on multiple contrast/challenge sets, even better than CAD

But only when we do targeted augmentation!

Model already knows woman ≠ man well, **need to focus on its error cases, e.g., preposition, quantifier!**

v.s. Manual creation: more effective

Kaushik et al. ICLR'20

Workers spent roughly 5 minutes per revised review, and 4 minutes per revised sentence (for NLI).

Gardner el al. EMNLP 20

30 seconds to change one image caption sentence.

Verify the machine generated ones, **cheaper than** writing new ones

Focus on perturbed parts, **cheaper than** parsing full sentences

→ **30s per round (3 perturbations)**

Counterfactual explanation & analysis

Do counterfactuals help with model understanding?

Counterfactual explanation: Complement *Status-quo*

Q1: How can I help a friend experiencing serious depression?

Q2: How do I help a friend who is in depression?

Predict $f(x)$: = *Duplicate* (98.2% confident)

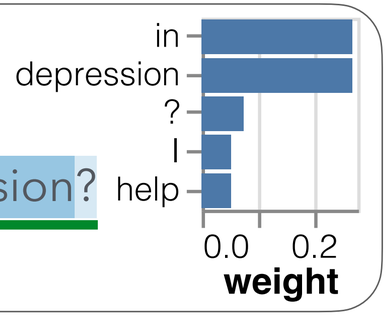
Counterfactual explanation: Complement *Status-quo*

Evaluate with popular saliency score (e.g. SHAP)

Q1: How can I help a friend experiencing serious depression?

Q2: How do I help a friend who is in depression?

Predict $f(x)$: = *Duplicate* (98.2% confident)



Counterfactual explanation: Complement *Status-quo*

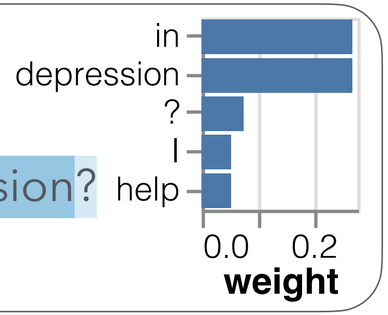
Evaluate with popular saliency score (e.g. SHAP)

Q1: How can I help a friend experiencing serious depression?

A

Q2: How do I help a friend who is in depression?

Predict $f(x)$: = *Duplicate* (98.2% confident)



But…

$\hat{x}$, perturbed Q2 $f(\hat{x})$

Q2: How do I find a friend who is in depression? =

✗ salience score is too abstract

Counterfactual explanation: Complement *Status-quo*

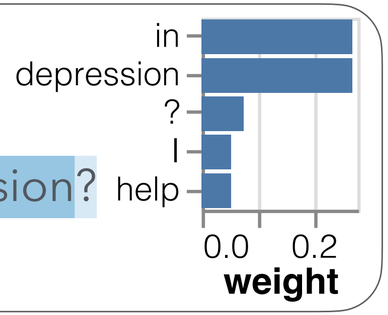
Evaluate with popular saliency score (e.g. SHAP)

Q1: How can I help a friend experiencing serious depression?

A

Q2: How do I help a friend who is in depression?

Predict $f(x)$: = *Duplicate* (98.2% confident)



But…

$\hat{x}$, perturbed Q2 $f(\hat{x})$

Q2: How do I find a friend who is in depression?	=
Q2: How do I help a woman who is in depression?	≠

× salience score is too abstract

× they miss error regions

Counterfactual explanation: Complement *Status-quo*

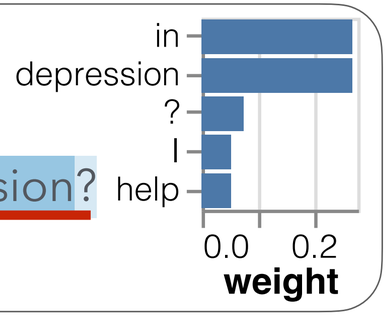
Evaluate with popular saliency score (e.g. SHAP)

Q1: How can I help a friend experiencing serious depression?

A

Q2: How do I help a friend who is in depression?

Predict $f(x)$: = *Duplicate* (98.2% confident)



But…

$\hat{x}$, perturbed Q2	$f(\hat{x})$
Q2: How do I find a friend who is in depression?	=
Q2: How do I help a woman who is in depression?	≠
Q2: How do I help a friend who is suicidal?	=

× salience score is too abstract

× they miss error regions

Counterfactual explanation: Complement *Status-quo*

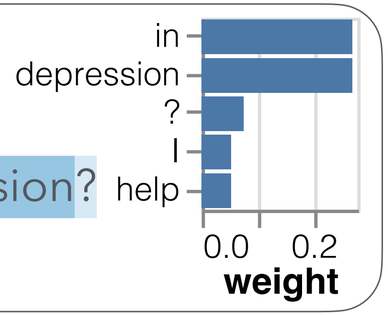
Evaluate with popular saliency score (e.g. SHAP)

Q1: How can I help a friend experiencing serious depression?

A

Q2: How do I help a friend who is in depression?

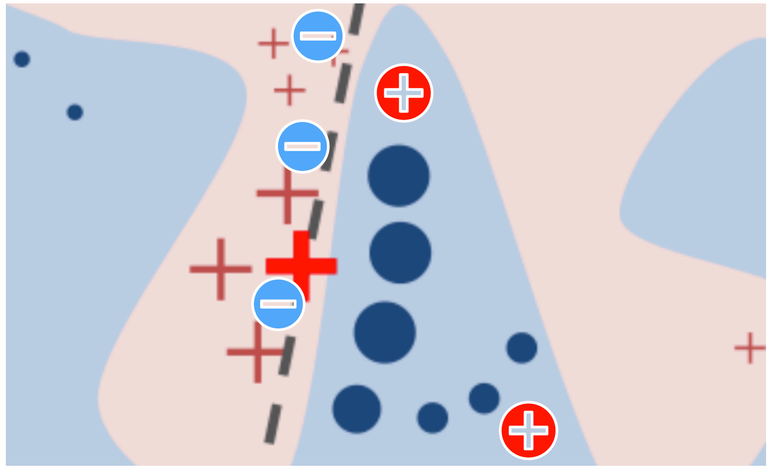
Predict $f(x)$: = *Duplicate* (98.2% confident)



But…

$\hat{x}$, perturbed Q2 $f(\hat{x})$

Q2: How do I •find a friend who is in depression?	=
Q2: How do I help a •woman who is in depression?	≠
Q2: How do I help a friend who is •suicidal?	=



SHAP: Estimate scores by masking

How <mask> help <mask> <mask>…

Not reflect model behavior on natural counterfactuals…

Complement SHAP with surprising (violate expectation) counterfactuals

Counterfactual explanation: Complement *Status-quo*

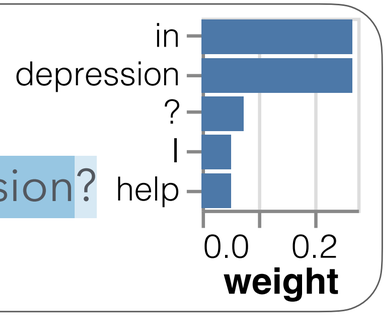
Evaluate with popular saliency score (e.g. SHAP)

Q1: How can I help a friend experiencing serious depression?

A

Q2: How do I help a friend who is in depression?

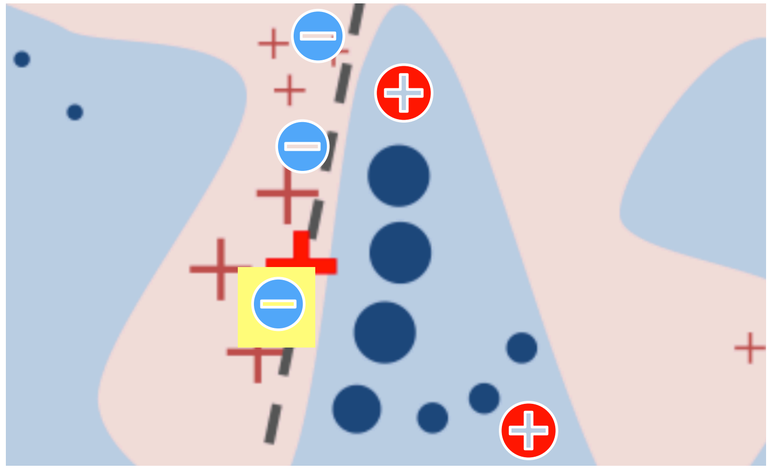
Predict $f(x)$: = *Duplicate* (98.2% confident)



But…

$\hat{x}$, perturbed Q2 $f(\hat{x})$

Q2: How do I •find a friend who is in depression?	=
Q2: How do I help a •woman who is in depression?	≠
Q2: How do I help a friend who is •suicidal?	=



SHAP: Estimate scores by masking

How <mask> help <mask> <mask>…

Not reflect model behavior on natural counterfactuals…

Complement SHAP with surprising (violate expectation) counterfactuals

Counterfactual explanation: Complement *Status-quo*

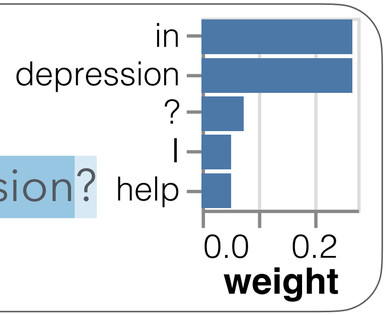
Evaluate with popular saliency score (e.g. SHAP)

Q1: How can I help a friend experiencing serious depression?

A

Q2: How do I help a friend who is in depression?

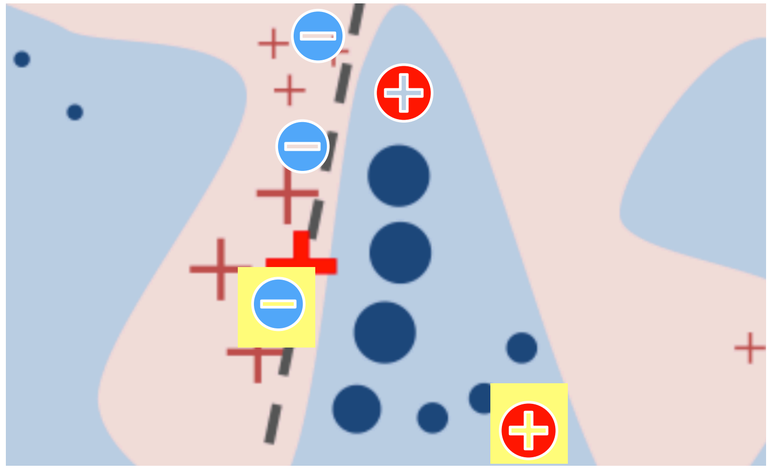
Predict $f(x)$: = *Duplicate* (98.2% confident)



But…

$\hat{x}$, perturbed Q2 $f(\hat{x})$

Q2: How do I •find a friend who is in depression?	=
Q2: How do I help a •woman who is in depression?	≠
Q2: How do I help a friend who is •suicidal?	=



SHAP: Estimate scores by masking

How <mask> help <mask> <mask>…

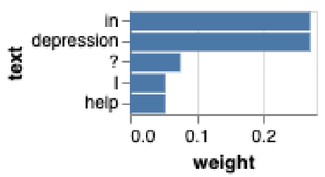
Not reflect model behavior on natural counterfactuals…

Complement SHAP with surprising (violate expectation) counterfactuals

User study: Surprising points bring additional insights

REFERENCE EXAMPLE AND WORD IMPORTANCE

Old Q1 How can I help a friend experiencing serious depression ?
Old Q2 How do I help a friend who is in depression ?
Model predicts Duplicate (97.3% confident)
Model correct? Correct



ASK QUESTIONS!

You can edit Q2 and get model's prediction on **variations** of this example. Use it wisely, and try to come up with edits that will best help you understand **the model behavior around** the instance!

Old Q1 How can I help a friend experiencing serious depression ?
New Q2 How do I play with a friend who is in depression ?

Get prediction

QUERY RESULTS WILL BE DISPLAYED BELOW

The green color highlights new words added in New Q2 , compared to Old Q2 in the **Reference example above.** • indicates something is deleted.

Old Q1 How can I help a friend experiencing serious depression ?
New Q2 How do I play with a friend who is in depression ?
Model predicts Non-duplicate (67.4% confident)

Old Q1 How can I help a friend experiencing serious depression ?
New Q2 How do I help a friend who is in frustrated ?
Model predicts Non-duplicate (95.1% confident)

GIVE Original examples, SHAP

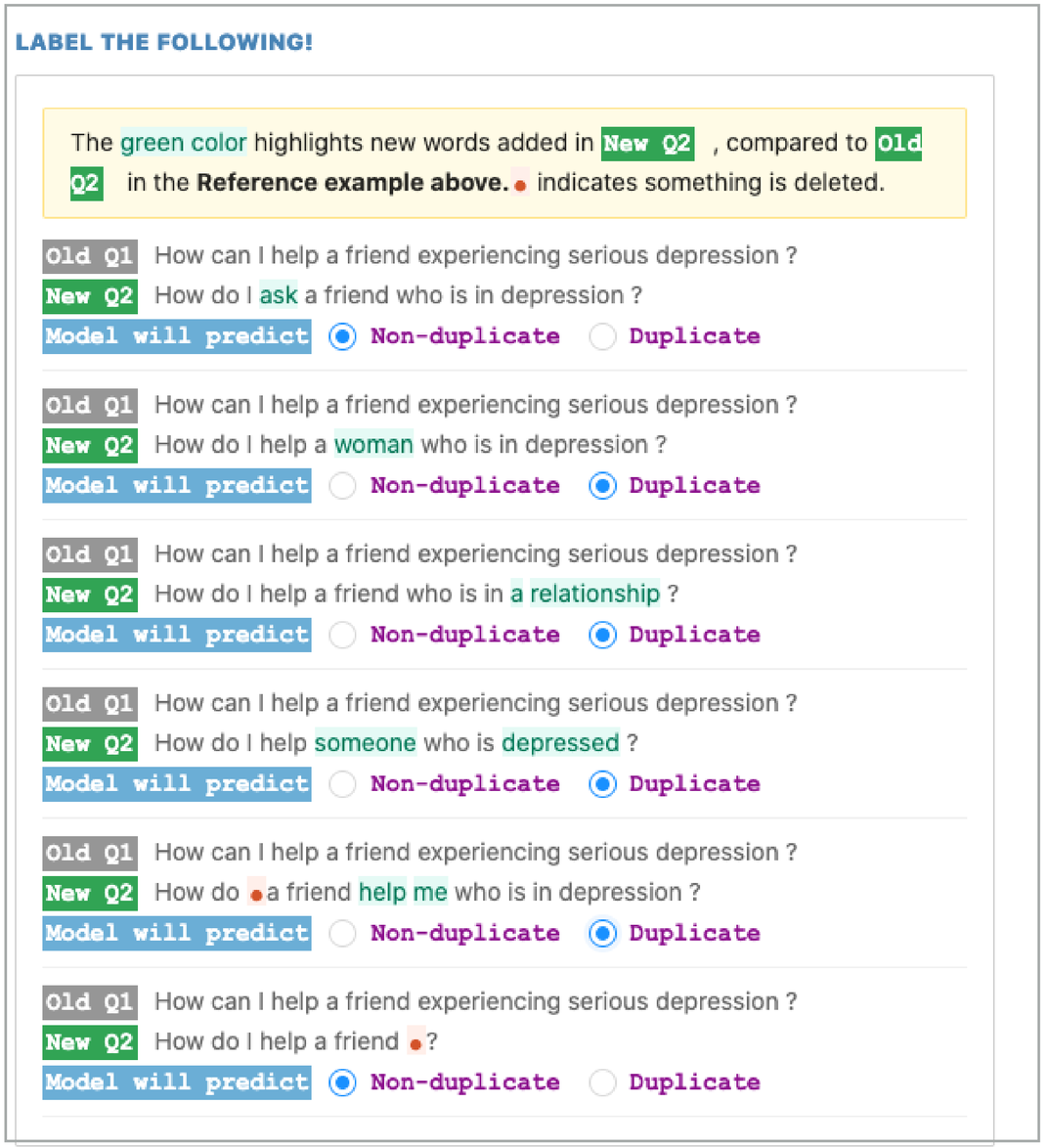
GIVE Manual counterfactual analysis

User study: Surprising points bring additional insights

GET Simulate model's behavior on Polyjuice surprising counterfactuals

13 NLP PhD students did only slightly better than random guessing. (57% correct)

these examples are surprising even after seeing explanations and creating manual counterfactuals!



Error analysis: Extend back to grouping

x	$f(x)$
P: A woman is holding a baby by a window. H: This woman is looking out the window.	*Neutral*
$\hat{x}$, perturbed H through [negation]	$f(\hat{x})$
H: No woman is looking out the window.	*Contradiction*
H: This woman isn't looking out the window.	*Contradiction*
H: This woman is not looking out the window.	*Neutral*

Neg correlates with Contradiction in NLI?

Has more nuances!

Inconsistency between "n't" and "not"!

$x \rightarrow f(\hat{x})$	Template	Coverage (%*N*→*C*)
…is not looking…	`AUX → AUX not`	
…aren't playing…	`* → * not`	412 (42.3%)
The→No girls like…	`* → * n't`	434 (43.5%)
	`* → * PART`	
A→No man in…	**`DET → No`**	180 (**92.8%**)

`DET → No` flips model prediction much more frequently!

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. Annotation Artifacts in Natural Language Inference Data. ACL 2020

Wu, T., Wongsuphasawat, K., Ren, D., Patel, K., & DuBois, C.. Tempura: Query Analysis with Structural Templates. CHI 2020

Building blocks

Polyjuice

Quantitative grouping

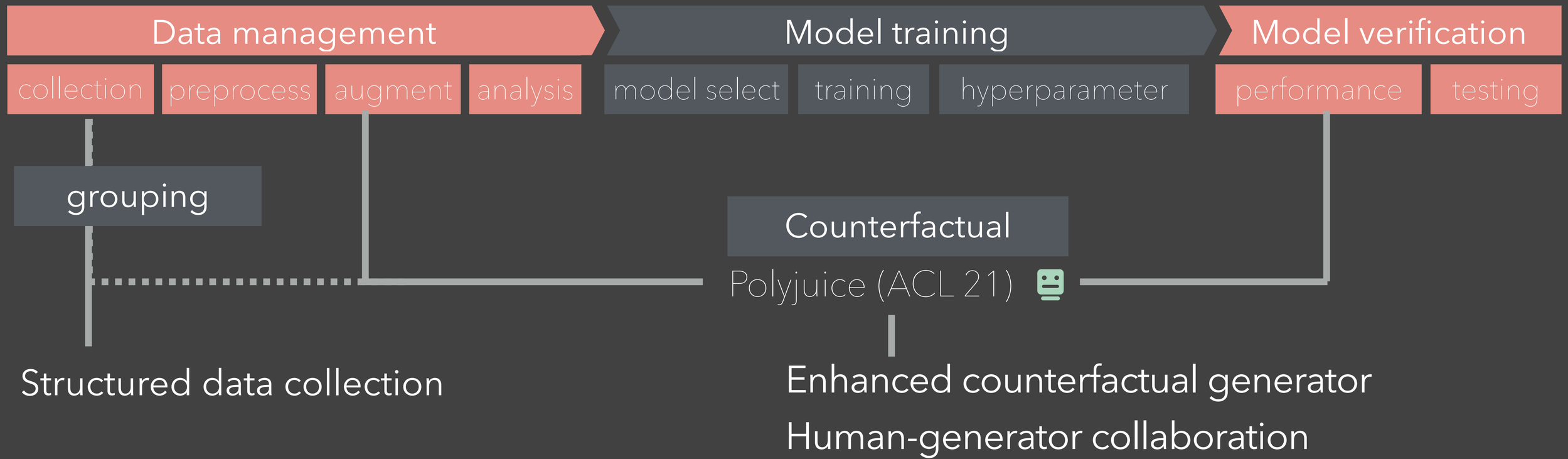
Inspect similar instances, semantically & syntactically

Supports

Counterfactual perturbation

Isolate important components targeted minimal rewrites

General-purpose generator with control, for various applications


Data management
Model training
Model verification
collection
preprocess
augment
analysis
model select
training
hyperparameter
performance
testing
grouping
Counterfactual
Polyjuice (ACL 21)
Structured data collection
Enhanced counterfactual generator
Human-generator collaboration

Polyjuice 2.0, with structured labels & content keywords

Polyjuice is nice, but not bias free.

× Control codes are predefined and not exhaustive

× Paired data makes certain perturbations more or less likely.

× Does not accept any context-control.

→ **Design generators that rely on**

Multiple blanks,

linguistic structural labels (SRL), for choosing which blank to fill in + rough content

content signal, for more concrete fill-in control

Polyjuice 2.0, with structured labels & content keywords

→Design generators that rely on

Multiple blanks,

linguistic structural labels (SRL), for choosing which blank to fill in + rough content

content signal, for more concrete fill-in control

MANNER	*
TEMPORAL	*
CAUSE	*
LOCATIVE	*

And `[BLANK]` he `[BLANK]` nearly always bought and sold `[BLANK]`.

`[EMTPY]`	`[EMTPY]`	like a good guy.
`[EMTPY]`	through a 17-month period	`[EMTPY]`
moreover because he had always known him ,		`[EMTPY]`
`[EMTPY]`	`[EMTPY]`	in the United States

Polyjuice 2.0, with structured labels & content keywords

→Design generators that rely on

Multiple blanks,

linguistic structural labels (SRL), for choosing which blank to fill in + rough content

content signal, for more concrete fill-in control

MANNER	*
TEMPORAL	*
CAUSE	*
LOCATIVE	*
LOCATIVE	profit

And `[BLANK]` he `[BLANK]` nearly always bought and sold `[BLANK]`.

`[EMTPY]`	`[EMTPY]`	like a good guy.
`[EMTPY]`	through a 17-month period	`[EMTPY]`
moreover because he had always known him ,		`[EMTPY]`
`[EMTPY]`	`[EMTPY]`	in the United States
`[EMTPY]`	`[EMTPY]`	at a profit

Interactive workflow: Human-generator interaction

Scalable

Possibly have more vocabulary

Can pick most/least impact changing places

Can be more creative after seeing the machine generation

saw is → isn't, can think of is → could have been

Humans assist

Machine generates/selects

Human only verify

Humans lead

Human provide seed perturbations

Human BLANKs change points

Machine follows up to broaden vocab.

Take turns

Machine starts

Human corrects imperfect ones

Human supplies missing cases

. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. CHI 2018.

Extend to labeling: Get structures at collection

Improve data collection, instead of apply structures “after the fact”

Infer the structure from labeled counterfactual pairs.

Labels + Has-negation

negative ↓ positive

I didn’t → did like the movie.

I don’t → do think it is a good book, and → even though it is expensive.

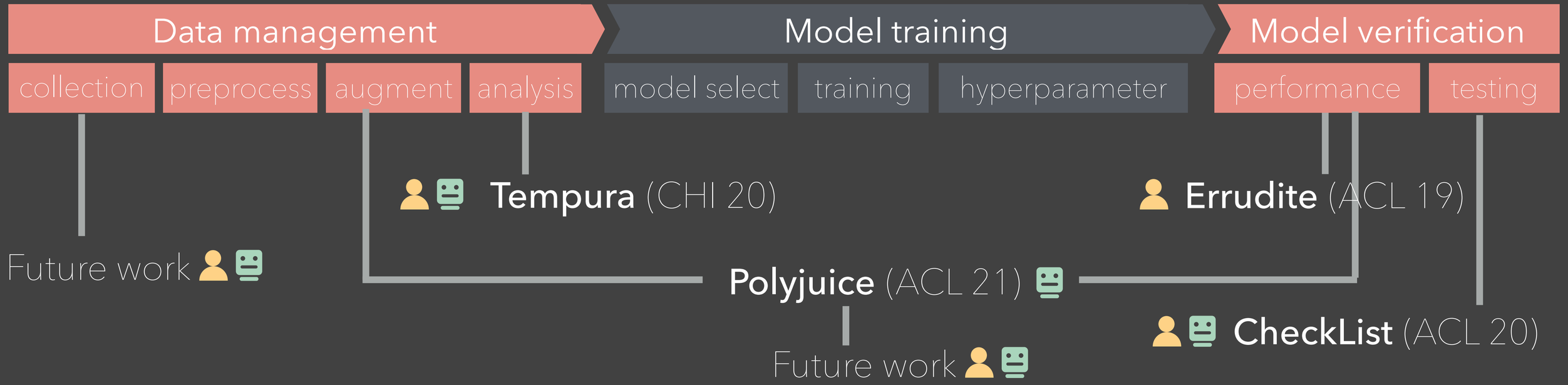
You would never → always care about this movie if you hadn’t seen the previous one.

Building blocks

Quantitative grouping

Counterfactual perturbation

Model development stages (Paleyes et al. 2020)

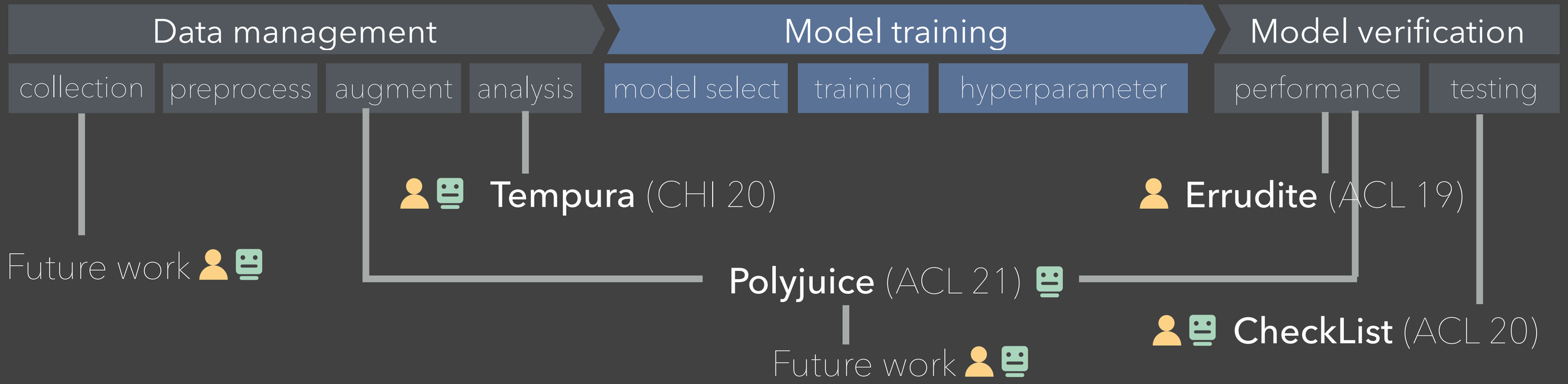


Building blocks

Quantitative grouping

Counterfactual perturbation

Model development stages (Paleyes et al. 2020)



Thanks!

And I’m happy to take questions :)

Backup slides

ACL 2019

Errudite: Scalable, Reproducible, and Testable Error Analysis

Tongshuang (Sherry) Wu @tongshuangwu
University of Washington

Marco Tulio Ribeiro
Microsoft Research

Jeffrey Heer @jeffrey_heer
Daniel S. Weld @dsweld
University of Washington




Deliveries: **Precise** + **Reproducible** + **Re-applicable**

Attribute

```
ENT(g)
```

Groups

```
all_instance
is_entity
has_distractor
correct_type
is_distracted
```

Rewrite rule

```
rewrite(
   c,
   string(p(m))→"#")
```

Deliveries: **Precise** + **Reproducible** + **Re-applicable**

Attribute **+** **Groups** **+** **Rewrite rule**

applied to…

BiDAF is …

not particularly **bad** at distractors.

Seemingly distractor errors can be due to **other factors**.

Deliveries: **Precise** + **Reproducible** + **Re-applicable**

Attribute + Groups + Rewrite rule

Re-applied to…

Other datasets & **Other** models…

? at handling distractor.

User study: **What is imprecise answer boundaries?**

"The model is making predictions with missing or additional words…?"

D1 No exact match, but high overlap

```
exact_match(p(m)) == 0
  and f1(p(m)) > 0.7
```

D2 Off by at most 2 tokens both on the left and right

```
exact_match(p(m)) == 0
  and abs(answer_offset(p(m),"left")) <= 2
  and abs(answer_offset(p(m),"right")) <= 2
```

User study: **What is imprecise answer boundaries?**

"The model is making predictions with missing or additional words…?"

D1 No exact match, but high overlap

```
exact_match(p(m)) == 0
  and f1(p(m)) > 0.7
```

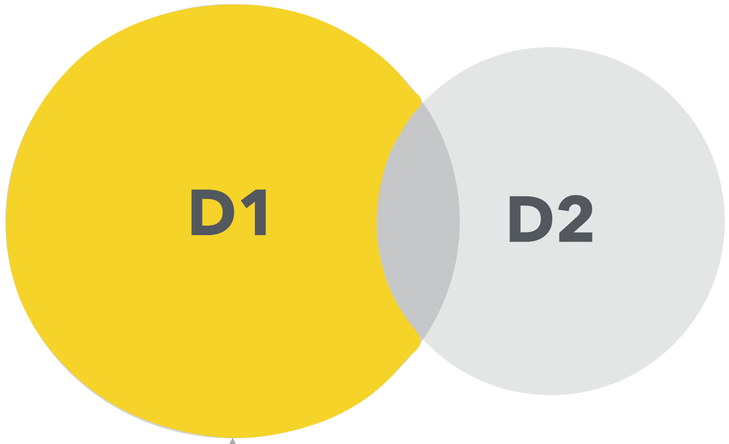
D2 Off by at most 2 tokens both on the left and right

```
exact_match(p(m)) == 0
  and abs(answer_offset(p(m),"left")) <= 2
  and abs(answer_offset(p(m),"right")) <= 2
```

User study: **What is imprecise answer boundaries?**

D1 No exact match, but high overlap

D2 Off by at most 2 tokens both on the left and right



`groundtruth`

…the **polynomial time hierarchy** collapses.

…believed that the polynomial hierarchy does..

`prediction`

ACL 2020 Best Paper Award

Beyond Accuracy: Behavioral Testing of NLP Models with **Checklist**

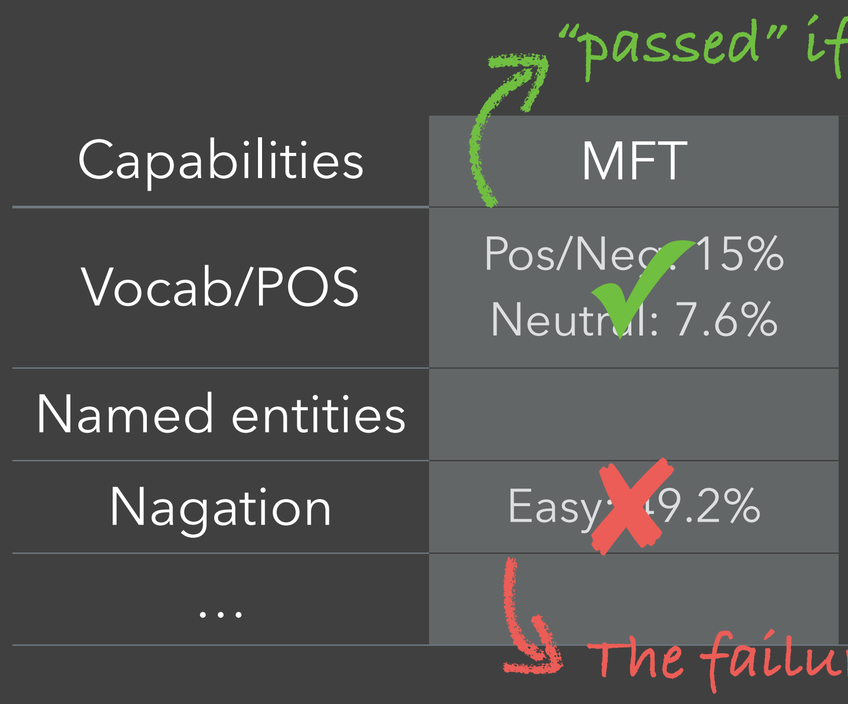
Marco Tulio Ribeiro
Microsoft Research

Tongshuang (Sherry) Wu @tongshuangwu
University of Washington

Jeffrey Heer @jeffrey_heer
Daniel S. Weld @dsweld
University of Washington




Discussion: translate failure rate to success / failure?



Affected by the test cases selected

Abs. value is not as interesting as "high enough"

Can be subjective & case-to-case

Discussion: Cautious on what to claim!

Failing a test ≠ failing what the test name indicates.

Linguistic capabilities are more intertwined. Should try to further isolate compounds through INV tests. And should fix the pattern anyways!

Passing a test ≠ model working.

Test cases are not comprehensive; Only give you more confident that the basic works.

Discussion: who are the users?

Model developers, Experts on model evaluation & task

Common and intuitive tests that are crucial for deployment

Researchers, Experts on model evaluation

Investigate into sophisticated tests (that may worth a paper)

Customers, Experts on the specific data/application

Tests specific to the dataset (e.g., NER tests on medical terms)

Ultimate goal: Have a shared test suite for each NLP task

User study: people test the same model/capability with different test cases!

Discussion: why do we care?

It's true ...

Some of the failures are by design and are not surprising.

e.g. MFT tests are usually out of distribution; SQuAD dataset do not have very short paragraphs.

But!

It is annotation artifact.

Dataset collection does not reflect the real world what we care about.

The training data will never be comprehensive.

Language is high dimension and selection bias is unavoidable.

The training data will keep getting more biased.

Concept drift caused by the deployed model interacting with the world.

Discussion: why do we care?

It's true ... The testing does not necessarily point to the source of bug / a fix.

NER-INV failure is due to contextual embedding, not my model/data.

But! We should first find the bug, and then try to isolate the source.

Detecting bugs is paramount for evaluation, and a prerequisite for further exploration of what caused them.

It's true ... Testing sophisticated capabilities can be hard.

Test cases for sarcasm require more effort than simple negation.

But! We can start with the simple ones as demo-ed!

Test models with the basics, & write tests close to models' capability.

Make sure your model pass level 1 MFTs before you reach level 3!

Data manage. | Model verification

augment | perform - explanation

Submitted to ACL 2021

Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving models

Tongshuang (Sherry) Wu @tongshuangwu
University of Washington

Marco Tulio Ribeiro
Microsoft Research

Jeffrey Heer @jeffrey_heer
Daniel S. Weld @dsweld
University of Washington




Have we selected criticizing points?

CLAIM

The “surprises” compensate SHAP.

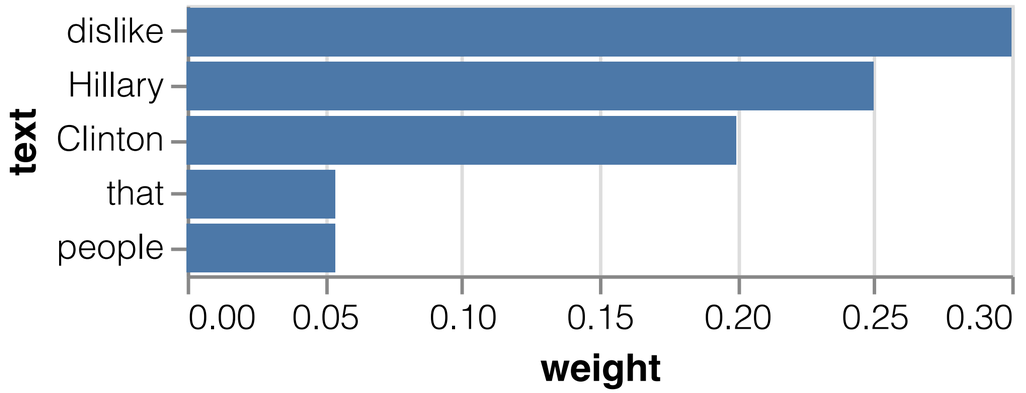
HOW? for each round… (20 in total)

GIVE Provide original examples, SHAP

Q1: Why do so many people hate Hilary Clinton?

Q2: What are the reasons that people dislike Hillary Clinton?

Predict: Duplicate (99.3% confident)



Have we selected criticizing points?

CLAIM

The “surprises” compensate SHAP.

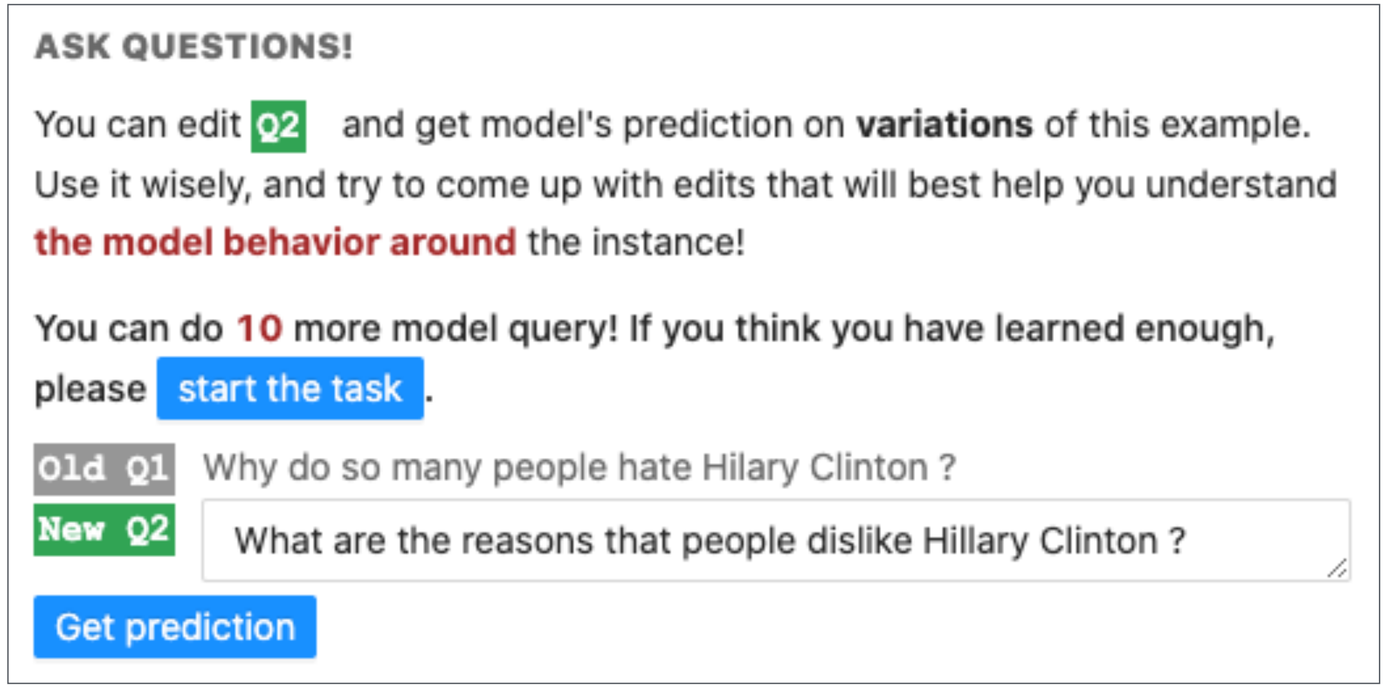
HOW? for each round… (20 in total)

GIVE Provide original examples, SHAP

GIVE Allow asking models questions

MEASURE:

unsuccessful simulation



Have we selected criticizing points?

CLAIM

The “surprises” compensate SHAP.

HOW? for each round… (20 in total)

GIVE Provide original examples, SHAP

GIVE Allow asking models questions

GET Simulate model’s behavior

on perturb. (criticism, random, human)

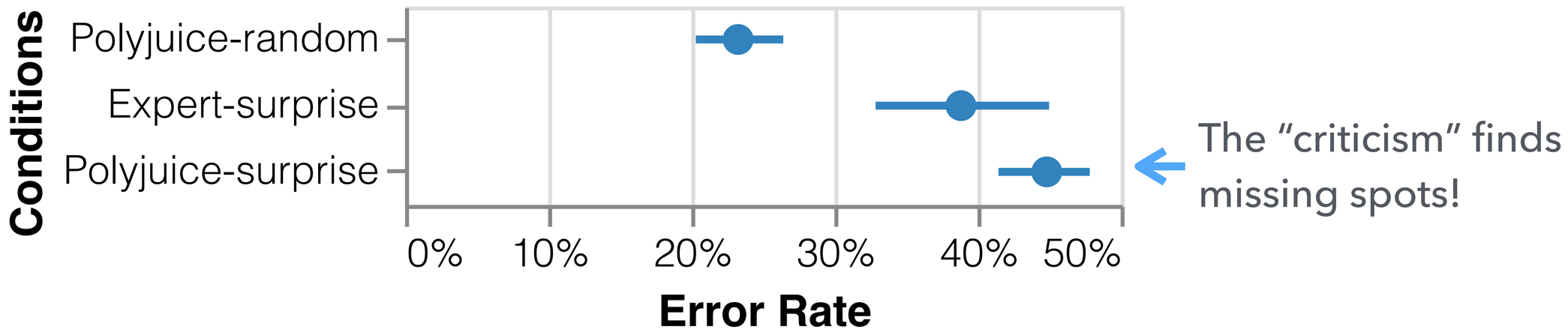
MEASURE:

Error rate: # unsuccessful simulation

The more they miss (V.S. random/human), the more information we add!

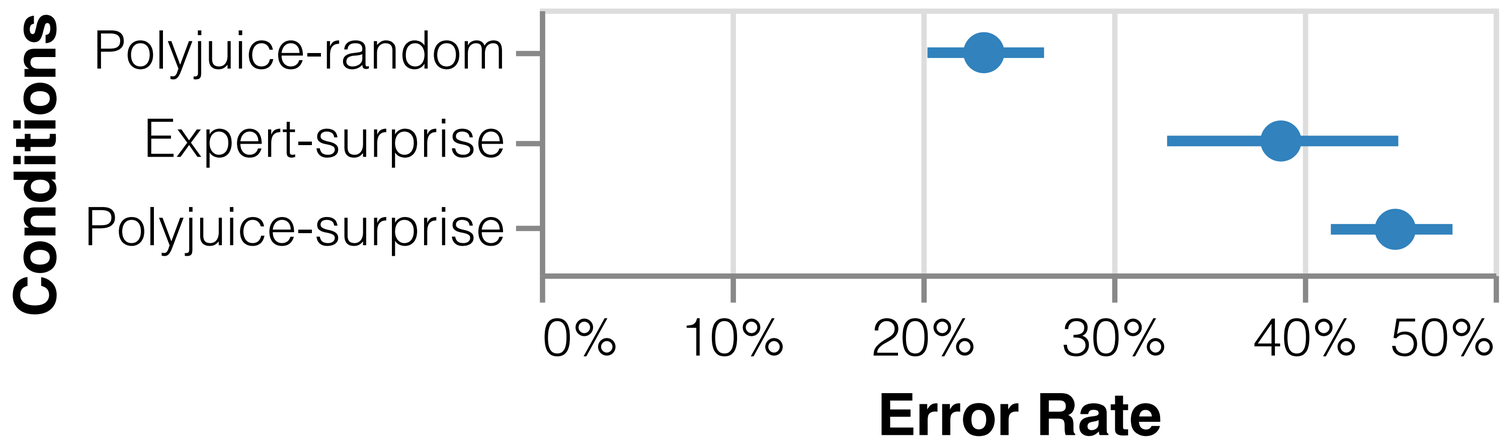
Results

13 people, 20 rounds, each round contains 2 questions × 3 selection conditions



Results

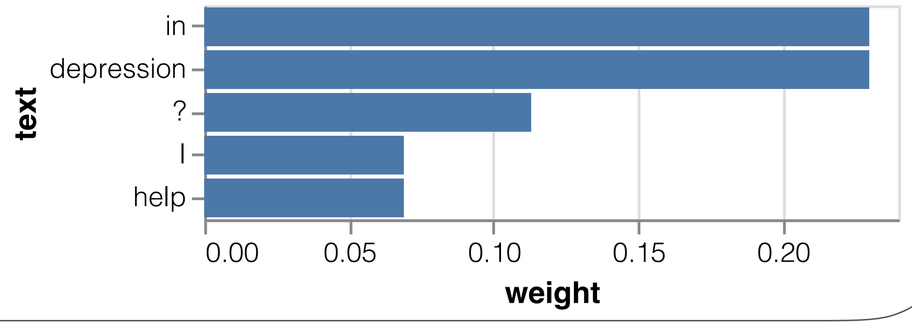
13 people, 20 rounds, each round contains 2 questions × 3 selection conditions



also find missing points,
but expensive!

Why did people miss the surprises?

Q1: How can I help a friend who is experiencing serious depression?
Q2: How do I help a friend who is in depression?
Predict: Duplicate (99.3% confident)



Surprise

Q1: How can I help a friend who is experiencing serious depression?
Q2: How do I help a friend → woman who is in depression?
Predict: Duplicate (99.3% confident)

Users usually query "important" words…

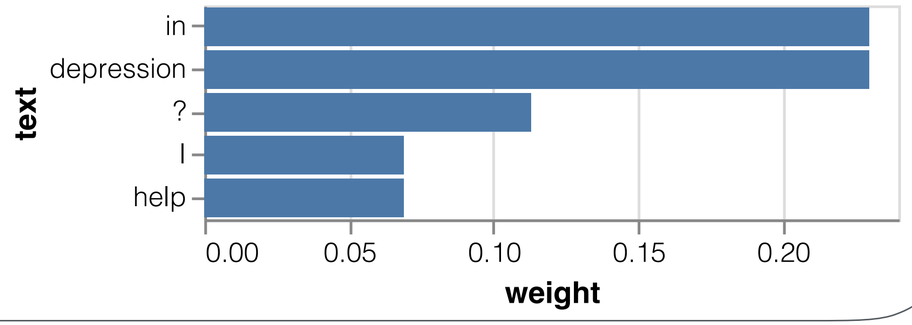
Q2: How do I help a friend who is in

depression → disease
depression → frustration
depression → happiness

85% of user inspections are on top features!

Why did people miss the surprises?

Q1: How can I help a friend who is experiencing serious depression?
Q2: How do I help a friend who is in depression?
Predict: Duplicate (99.3% confident)



Surprise

Q1: How can I help a friend who is experiencing serious depression?
Q2: How do I help → find a friend who is in depression?
Predict: Duplicate (99.3% confident)

Users overfit to their inspection…

Q1: How can I help a friend who is experiencing serious depression?
Q2: How do I help → play with a friend who is in depression?
Predict: Non-duplicate (99.3% confident)