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

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

Problem

Accurate NLP models still have blind spots or lacking capabilities.
Status quo process can be ad-hoc and biased.

Goal

To help practitioners gain systematic insights into their models' behaviors,
by organizing and exploring the inputs and outputs of their models.

Methods

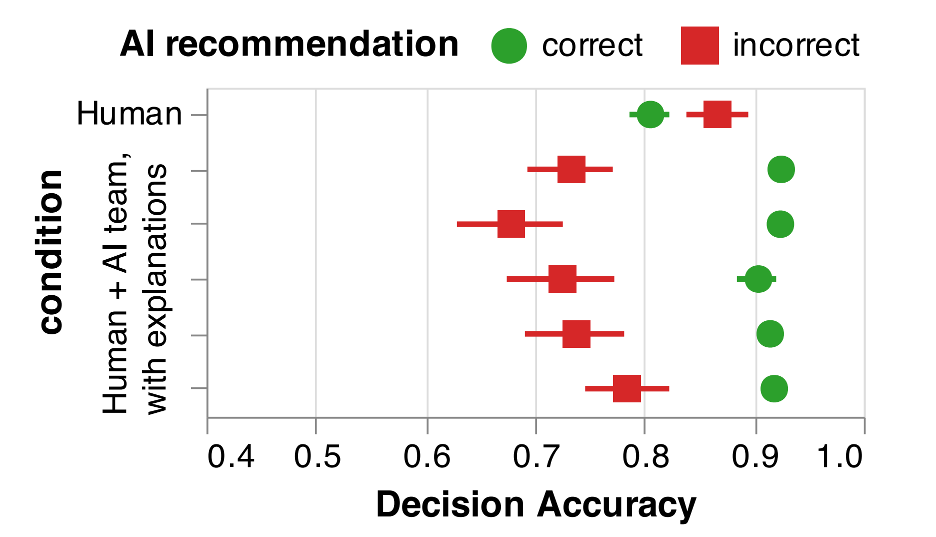
User studies to explore issues in the status-quo process;
Design principles and interactive tools to assist systematic analysis.

Q1: Pitfalls in status-quo local understanding/interventions?

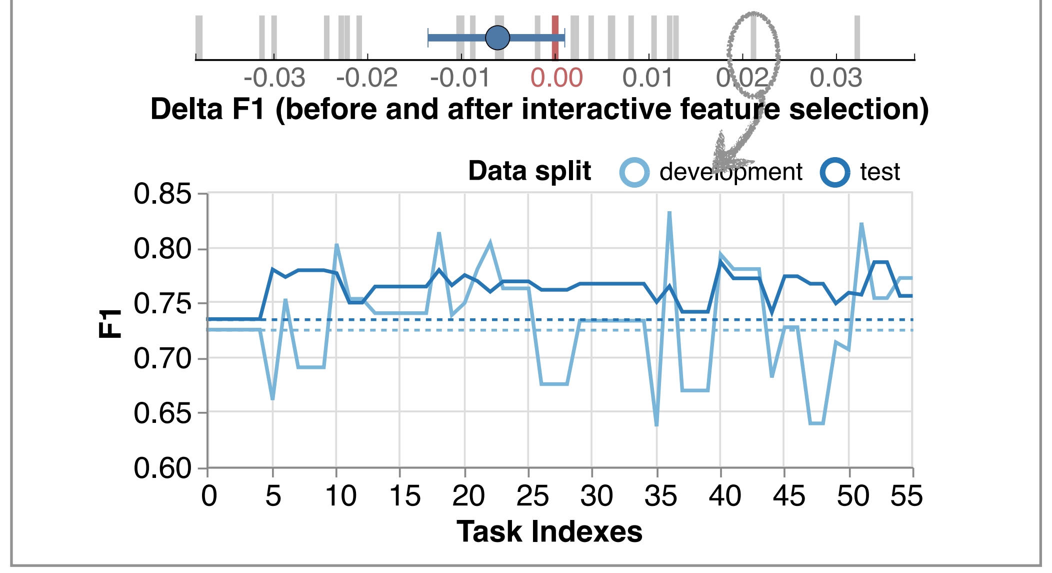
DID User studies to explore *status-quo* model understanding and updating.

TL;DR Users may "overfit" to local phenomena. Promote systematic understanding – error analysis, data quality assessment, augmentation.

People blindly follow (wrong) local explanations.



Interactive refinement makes F1oscillate, not improve.



Q2: What components are essential for systematic analyses?

DID

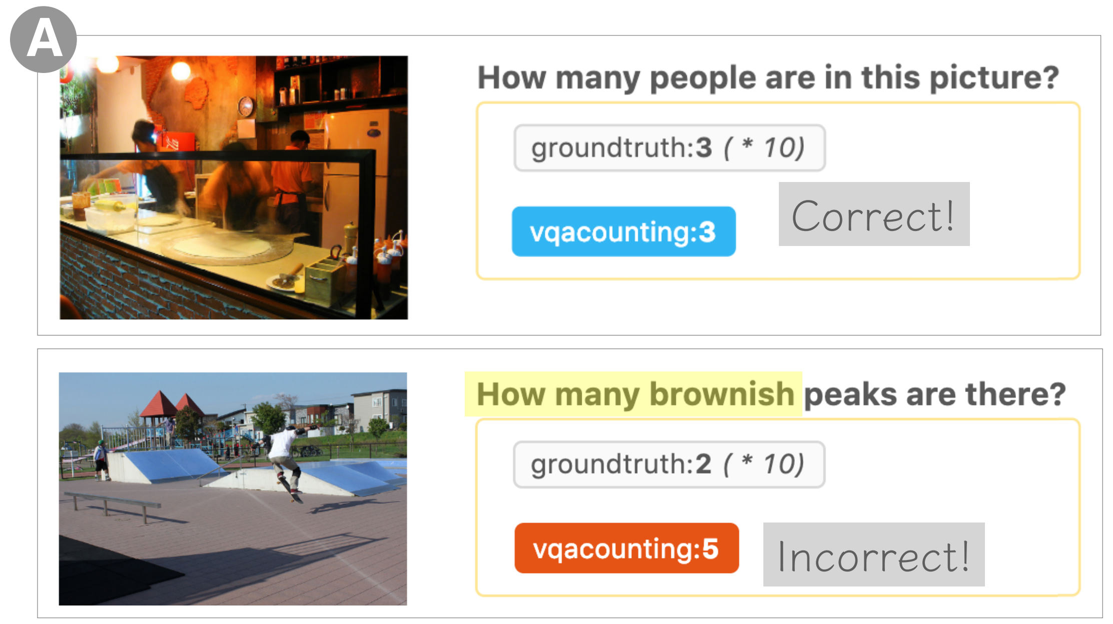
Identified building blocks

Built *Errudite*, error analysis tool with the building blocks

TL;DR

1. Group similar instances to scale the analysis
2. Counterfactual perturbations to isolate components

Prior error analysis process are not reproducible!



How many people are in this picture?

groundtruth:3 (* 10)

vqacounting:3

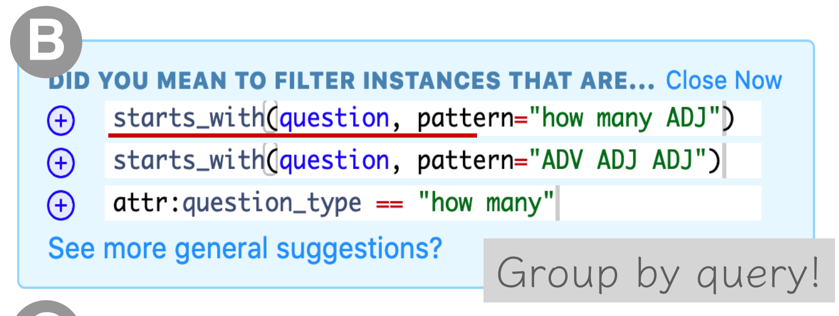
Correct!

How many brownish peaks are there?

groundtruth:2 (* 10)

vqacounting:5

Incorrect!

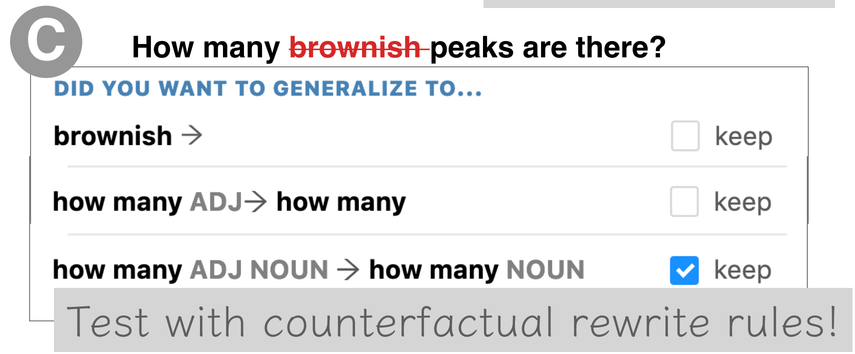


DID YOU MEAN TO FILTER INSTANCES THAT ARE... Close Now

```
⊕ starts_with(question, pattern="how many ADJ")
⊕ starts_with(question, pattern="ADV ADJ ADJ")
⊕ attr:question_type == "how many"
```

See more general suggestions?

Group by query!



How many ~~brownish~~ peaks are there?

DID YOU WANT TO GENERALIZE TO...

brownish → □ keep

how many ADJ→ how many □ keep

how many ADJ NOUN → how many NOUN ☑ keep

Test with counterfactual rewrite rules!

Q3: Grouping/counterfactual in different stages?

DID Designed grouping and perturbation methods & tools that prioritizes the inherent properties of tasks and datasets.

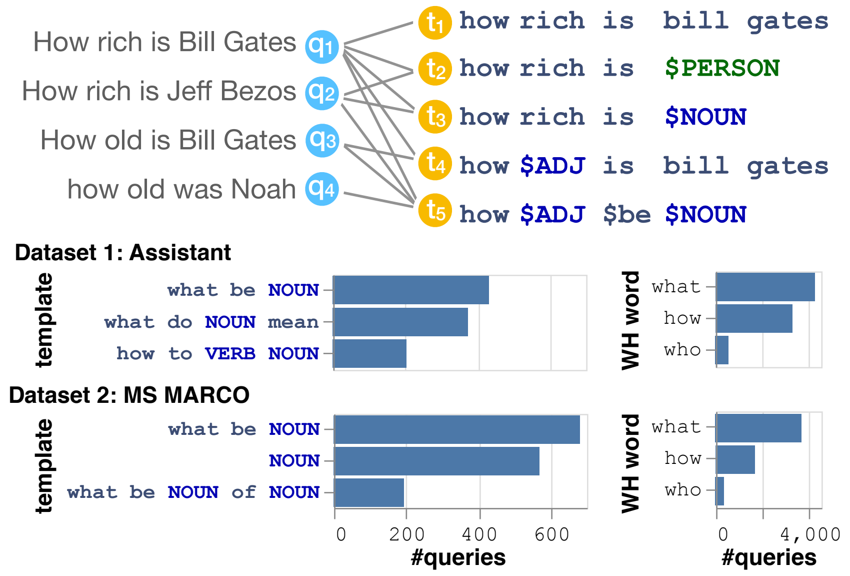
TL;DR

Tempura

What? Training data assessment

How? Mine and rank text group by linguistic based-templates.

So? Characterize dataset without manual filters (prone to biases).

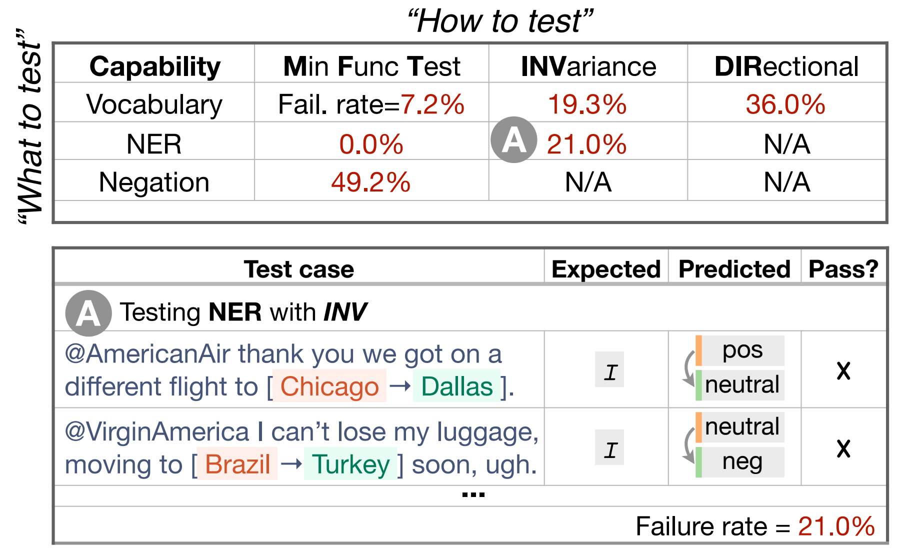


CheckList

What? Model testing framework

How? Guiding matrix for designing test case groups/perturbations.

So? CheckList users found 3 times as many bugs v.s. without it.


"How to test"
"What to test"

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=7.2%	19.3%	36.0%
NER	0.0%	A 21.0%	N/A
Negation	49.2%	N/A	N/A

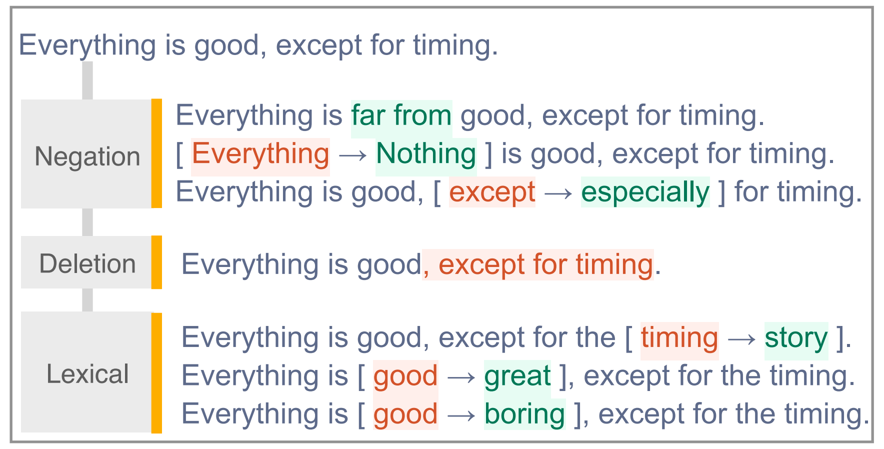
Test case	Expected	Predicted	Pass?
A Testing **NER** with ***INV***			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	*I*	pos neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	*I*	neutral neg	X
...			
Failure rate = 21.0%			

Aditor (ongoing)

What? Assisted perturbation

How? language models initialize rewrites + human correction.

So? counterfactual explanation, diverse contrast set ,



Everything is good, except for timing.

Negation
- Everything is far from good, except for timing.
- [Everything → Nothing] is good, except for timing.
- Everything is good, [except → especially] for timing.

Deletion
- Everything is good, except for timing.

Lexical
- Everything is good, except for the [timing → story].
- Everything is [good → great], except for the timing.
- Everything is [good → boring], except for the timing.

Takeaways

Local systematic analysis matters!

Grouping + counterfactuals are useful.

Different analysis stages, different tool designs.

Future work: broader applications

Analyze → collect data with structures

Experts → non-experts (e.g. hired annotators)

Structured labeling through grouping.

Get rationales through perturbations.

Evaluate → improve models

Data augmentation through perturbation.

Infer labeling functions.

References

Tongshuang Wu et al. Local decision pitfalls in interactive machine learning: An investigation into feature selection in sentiment analysis. TOCHI'19.

Tongshuang Wu et al. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. arXiv'20.

Tongshuang Wu, et al. Errudite: Scalable, reproducible, and testable error analysis. ACL'19

Tongshuang Wu, et al. Tempura: Query Analysis with Structural Templates. CHI'20.

Marco Tulio Ribeiro, Tongshuang Wu, et al. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. ACL'20.