Errudite: Scalable, Reproducible, and Testable Error Analysis

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, Daniel S. Weld
{ wtshuang, jheer, weld } @cs.washington.edu, marcotcr@microsoft.com

UNIVERSITY of WASHINGTON

Microsoft

Status Quo & Contribution

"We randomly select 50-100 incorrect questions (based on EM) and roughly label them into N error groups."

Biased conclusion due to…	Errudite
Subjective error groups	Precise & reproducible groups
+	+
Small samples	Scale up to the entire dev set
+	+
Focus exclusively on errors	Cover errors & correct instances
+	+
No test on true cause	Test via counterfactual analysis

Principle I: Be Precise

Precise Domain Specific Language: Quantify instances.

Qualitative Description: The model is bad on long questions

Quantitative Description: questions with more than 20 tokens

```
length(question) > 20
```

Attribute Extractor	Target	Operators
length	question	>
question_type	groundtruth	!=
answer_type	prediction (model)	in
	token	has_any

User study: What is imprecise answer boundaries?

"The model makes 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
```

Groundtruth
…**polynomial time hierarchy** collapses
…believed the polynomial hierarchy does…
Prediction

D1 D2

DSL reveals otherwise unnoticed definition diff!

Principle II: Cover All Data

An example ERROR
Task: Question Answering
Model: BiDAF

Q: Who created the 2005 theme for Doctor Who?
ENT(prediction) == "PERSON"
C: ...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... featured sampled from the 1963 original.
ENT(groundtruth) == "PERSON"

Generalize to Distractor Hypothesis: BiDAF matches entity types, but misses the exact answer.

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

			Incorrect	Correct
all_instances	N/A	10570	32%	68%
is_entity	Line 1	4240	20%	80%
has_distractor	Lines 1-3	3495	21%	79%
correct_type	Lines 1-4	2988	12%	88%
is_distracted	Lines 1-5	192	100%	

88% EM > 68% EM: BiDAF performs better when have distractors & entity type is matched, than overall, reject / revise the hypothesis!

5.7% of all BiDAF errors: The distractor hypothesis seems correct!

BiDAF is not particularly bad at distractors.

Principle III: Test the Error Hypothesis

Are the 192 instances really wrong due to the distractor? `rewrite(c, STRING(p(m)) → "#")`

29% Incorrect → Incorrect
Another distractor is still confusing the model!

Q: Who created the 2005 theme for Doctor Who?
C: ...~~John Debney #~~ created a new arrangement of Ron Grainer's ... **Murray Gold** provided a new arrangement...

48% Incorrect → Correct
The distractor was fooling the model!

Q: How many kilometers is Warsaw from the Carpathian Mountains?
C: Warsaw lies in east-central Poland about **300** km (~~190 #~~mi) from...

23% Unchanged
Other factors are at play!

Q: How many of Jacksonville's city residents are younger than 18?
C: ... with **23.9%** under the age of 18, ~~10.5%~~ # from 18 to 24...

Seemingly distractor errors can be due to other factors.

Deliveries

Attribute `ENT(g)`
+
Groups `is_entity` `has_distractor` `correct_type` `is_distracted`
+
Rewrite `rewrite(c, STRING(p(m))→"#")`

↓ Applied to…

BiDAF is …

not particularly bad at distractors.

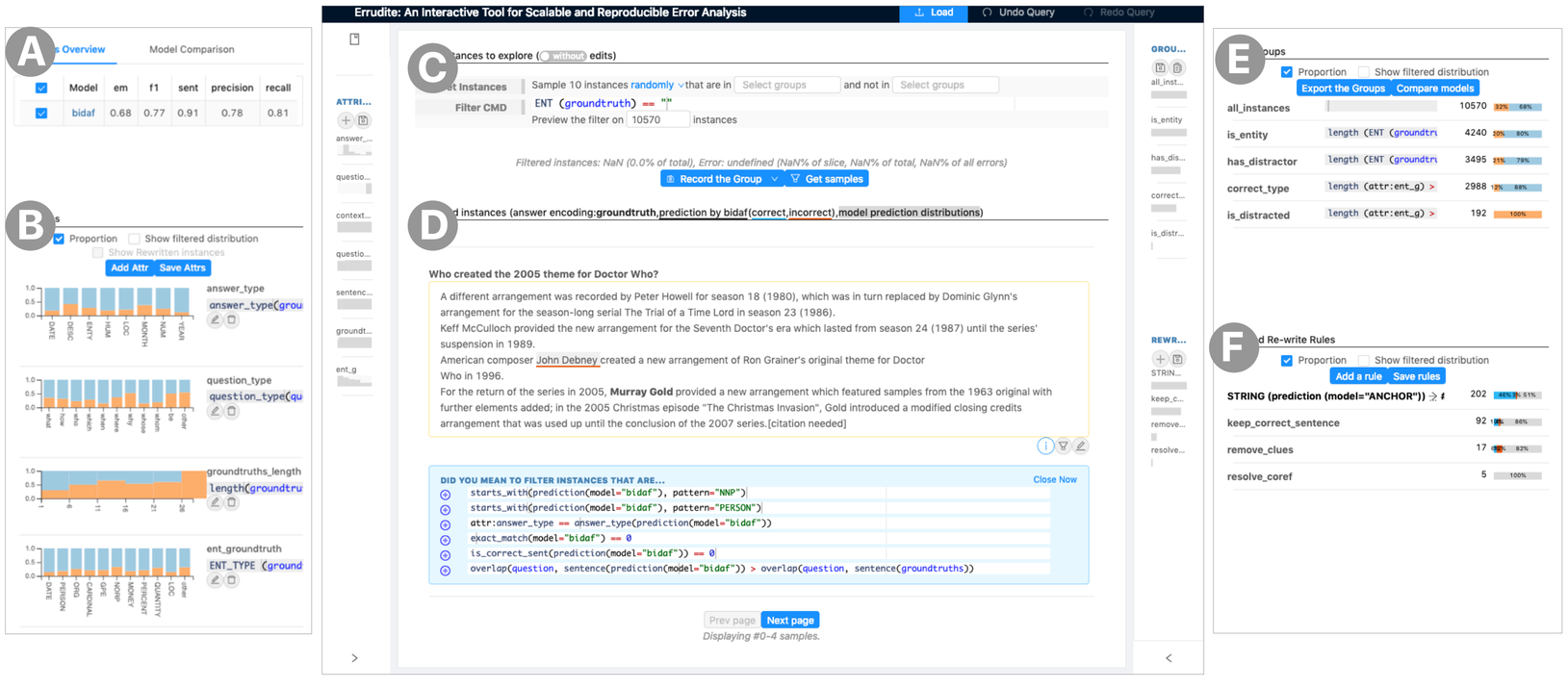
Seemingly distractor errors can be due to other factors.

↓ Re-apply to…

Other datasets & Other models…

? at handling distractor.

Resources



Demo Video

Opensource