RE for AI: What is an RS for an AI?

Daniel M. Berry
University of Waterloo

AIs and LMs

In the following,

an *artificial intelligence* (AI) could be …

a *learned machine* (LM) that results from …

teaching a *machine learning* (ML) app with …

some data from the *real world* (RW).

Tough Question of AIers

When I have asked AI people:

“How does one write a requirements specification (RS), S,

for an AI, A, for a task, T,

in a way that S can be used to decide whether

A correctly implements T,

by asking whether

A satisfies S?**”, ...**

Their Answer

they shrug their shoulders,

“Wer weis?”

Substitute Question

So, I ask:

“When you have build an AI, how do you know that it is correct?”

They go on and on about

- **recall, sensitivity**
- **precision**
- **specificity**
- **accuracy,**
- ***F*-measure,**
- **aut cetera**

Follow Up Question

I say:

“But those are evaluations! How do you know what measure values signal that the AI is correct?”

Again, I get a shoulder shrug.

Maybe what is missing are criteria that the measure values must satisfy for the AI to be accepted as being correct?

I Remember a Conversation

I remember an e-mail or chat exchange at an RE'21 workshop with Alessio Ferrari about reqs for AI.

He remarked very simply that …

an AI must mimic a human doing the same task.

So maybe the criteria must give measure values that prove that the AI mimics a human or is better.

A Recent Publication in EMSE

Then, in 2021, I published in *EMSE*, a paper explaining how to evaluate tools for hairy RE tasks that require NLP.

A *hairy* task is one that is difficult for humans to do well in the large scale, in the real world.

Its Advice

Its advice for evaluation the goodness of a tool for a hairy RE task includes:

- **The tool's**

  achieving high recall

  is usually an order of magnitude more important than

  achieving high precision.

Its Advice, Cont'd

- **The tool's *effective recall* after a human's manual vetting the tool's output**

  must be compared to

  a human's recall doing the task manually.

Its Advice, Cont'd

- **The time to run the tool and do the manual vetting**

  must be compared to

  the time for a human to do the entire task manually.

To Evaluate a Tool

Thus, the evaluation of a tool for a hairy task requires …

gathering more data than …

just its recall and precision …

against a gold set (ground truth).

To Evaluate a Tool, Cont'd

It requires gathering at least also the

- **average recall of humans against the same gold set,**
- **time for humans to find a true positive in building the gold set, and**
- **time for humans to reject a false positive from the tool.**

Main Insight of REFSQ Paper

The main insight of my REFSQ'22 paper is that a specification for an AI for a hairy task consists of

1. **a set of measures used for evaluation,**
2. **criteria that the measures must satisfy, and**
3. **other data about the context of the use of the AI, including the RW data that teaches an LM.**

Set of Measures

The set of measures used for evaluation measures *correctness* in some sense and is usually calculated from a confusion matrix, e.g.,

- **recall and precision,**
- **sensitivity and specificity,**
- **F-measure**
- **accuracy**

Criteria For Measures

The criteria that the measures must satisfy …

help show that the AI …

can be considered as …

mimicking or doing better than a human doing the same task.

Criteria, Cont'd

These criteria will usually include …

the values of the measures that humans *actually* achieve …

when doing the same task.

Other Data

The other data are data about the context of the use of the AI that …

- **allow engineering tradeoffs to help the AI meet the criteria and**

- **decide borderline cases.**

Allowing Engineering Tradeoffs

Example of allowing engineering tradeoffs to help the AI meet the criteria:

There is time for human vetting, e.g.,
in diagnosing radiographs for cancer OR
in recognizing stop signs to build teaching set for an AI
→ want recall = 100%, but low precision is fine

vs

real-time use, recognizing stop signs in AV
→ want recall = precision = 100%

Deciding Borderline Cases

Example of deciding borderline cases:

fast to vet,

beats human precision, *but*

tool recall is ε < human recall

For Details

For details, read the REFSQ'22 paper.

For details missing in the paper, google for a like-titled tech report at my Web site:

https://cs.uwaterloo.ca/~dberry/FTP_SITE/tech.reports/RE4AI_TechReport.pdf