Empirical Evaluation of Tools for Hairy Requirements Engineering Tasks

(Why You Should Bother to Read the Paper!)

Daniel M. Berry
University of Waterloo, dberry@uwaterloo.ca

You Just Developed a Tool

You have developed a tool, *T*, to find ambiguities in any input NL RE document.

And now you want to evaluate *T*.

Prepare to Evaluate

You have gotten a good-sized representative natural-language requirements specification, *I* (Input), on which to test *T*.

You have gotten a group of experts in the domain of *I* to construct a gold set, *G*, of the ambiguities in *I*.

Each expert constructed er own gold set, *g*,

and then the experts arrived at *G* by consensus.

Apply Tool to Test

You have run *T* on on *I*.

You have compared *T*’s output, *O*, with *G*.

You have determined from the comparison that *T* has

- **85% recall (how much of *G* was in *O*)**
- **40% precision (how much of *O* was in *G*)**

as an ambiguity finder.

The Question

The question is:

> **Is *T* a good tool for finding ambiguities in NL RE documents?**

Typical Answer

A typical answer is like:

"Well … 85% is not a bad recall, ….

***But* the recall is significantly less than 100%.**

***And* precision is kinda low!**

Typical Answer, Cont'd

Thus, a human will have to manually vet *O*

to weed out the false positives,

and that's a dull, boring, tedious job — yechh!

Not to mention that vetting itself will probably lose some of *O*'s true positive ambibuities from *O*."

☹

Reality

The reality is that we have no idea how good *T* is if we have only these data.

I mean:

> **From where did I get this idea that 85% is not a bad recall?**
>
> **Why is *any* imprecision bad if vetting is faster than a manual search of the original document?**

To Really Know

To really know how good *T*'s recall is, you need to know

how well *humans* do on *T*'s task of finding ambiguities in any input NL RE document.

If humans achieve only 70% recall, *T* is good.

If humans achieve 100% recall, *T* is bad.

Not So Simple

However, for humans to achieve 100% recall is unlikely.

Humans, including I, believe it or not, *do* make mistakes. ☺

So you need to know how much recall humans actually *do* achieve, the humanly achievable recall (HAR) of *T*'s task!

Recall?

But whoa!!!

Why this emphasis on recall …

(to the apparent exclusion of precision)?

Reason for Tool

If the problem for humans were not the difficulty of *finding ambiguities*,

even if the difficulty is one of only fatigue,

we would not be bothering to build *T* in the first place.

So the emphasis is on achieving high recall, at least better than humans can.

But Whoa!

But whoa!!!

What about the precision of 40%?

Gotta Vet Tool's Output

As mentioned ealier,

this low precision means that humans will have to vet *O* …

to weed out the false positives,

and these false positives make up 60% of the output. ☹

Right Data to the Rescue

Well …

IF it so happens that

> **the effective recall of *T* after manual vetting of *O* is 83%, and this 83% is higher than the HAR, and**
>
> **the time to manually vet *O* for false positives is less than the time to manually search *I* for true positives**

Data to the Rescue, Cont’d

THEN

T is good in spite of the low precision,

because …

Because

Ultimately,

running *T* followed by a human's vetting *O*

gets higher recall in less time than

than a human's searching *I* manually.

Underlying Assumption

We are talking about a tool for a task that *has* to be done!

There's *no* option *not* to do the task.

Tool’s Context

The context demands that the task be done.

E.g, quality control, safety, security, reliability, correctness, regulations, velc. demands it.

In such a context, …

the alternative to using a tool is doing the task manually!

This is the primary motivation for *building T* in the first place.

Implication

So *all* evaluations of any tool must be in comparison to how humans do the same task.

After all, with *no* option *not* to do the task, …

comparison of running the tool and vetting to the manual task is a fair comparison.

The J1st Paper

The paper describes how to

organize the usual experiment

that evaluates the recall and precision of *T*

so as to collect all the data you will need

to do a full evaluation that takes into account

humanly achievable recall (HAR) and
the context of *T*'s use.

Raw Data

These are the new data to gather during construction of *G*, i.e., new besides what are already gathered normally.

From each domain expert helping to construct *G*:

1. **er own set of ambiguities *g* (to get er own HAR), and**
2. **the time E spent to build *g* (to time er manual search).**

Raw Data, Cont'd

These are the new data to gather during the evaluation of *T* with *I* and *G*, i.e., new besides what are already gathered normally.

Necessarily only *estimates* of

1. **expected cost of failure to find a true positive, and**
2. **expected cost of accumulated nuisance of vetter's encountering *yet* another false positive.**

Raw Data, Cont’d

These are the new data to gather during the vetting *O* against *G*, i.e., new besides what are already gathered normally.

From each domain expert helping to vet:

1. **er own set of ambiguities in *O* (to get er own effective recall), and**
2. **the time E spent to vet *O* (to time er own vetting).**

Defensive Data Gathering

You may not need all of these for any evaluation, but

all of them must be gathered when they are available, because

they cannot be constructed later.

To Calculate

Here are the data to calculate for an evaluation, besides the standard recall, precision, and *F*-measure of *T*

1. **the HAR for *T*'s task (avg. of individual HARs),**
2. **the time to manually decide if an item in *I* is a true positive (TP),**
3. **the time to find a true positive manually in *I* (search *n* items to find one TP $\rightarrow$ *n*$\times$#2),**

To Calculate, Cont'd

4. **the time to vet an item in *O*,**
5. **effective recall after vetting (average of individual effective recalls),**
6. **summarization of *T*, the percentage of *I* that is eliminated from the human vetter's search in the tool's producing *O*, and**
7. **ratio of 2 to 4, vetting speed up per item (vetting is usually faster than manual search per item).**

Rational Evaluation

If you gather and calculate all the prescribed data, …

you will be able to do a rational evaluation of *T*

against human capabilities

in the context in which *T* will be applied.

Engineering Tradeoffs

You will be able to use these data to engineer tradeoffs,

e.g., between recall and precision.

Example of Engineering

Suppose you already have an algorithm for *T* whose effective recall is above the HAR.

You try a new algorthm with

- **better raw recall but**
- **worse precision and**
- **worse but still decent summarization.**

Is it worth using the new algorithm for *T*?

Maybe?

Or maybe not!

You carry out the evaluation and discover that

the greater imprecision and
the decrease in summarization

cause

vetting to be less accurate,

leading to a reduction in effective recall to below the HAR. ☹

Empirical Study

Paper treats a test of *T* as an empirical study and

talks about

- **confidence in results**
- **dealing with threats to the validity of conclusions,**
  - **including representativeness of *I*.**

RT_P

Now go read the ____ paper! ☺

Berry, D.M.
Empirical Evaluation of Tools for Hairy Requirements Engineering Tasks
Empirical Software Engineering
26:111, pp. 1–77, 2021
https://doi.org/10.1007/s10664-021-09986-0

Origin of This Work

I remember many times looking at a typical paper about

a new NLP-based tool for

a hairy (not conceptually hard, but unmanageable for real documents) RE task,

e.g., searching for abstractions, ambiguities, links, autc.,

My Interest

I was interested in the paper because

I knew that we needed the tool because

people were not good at the task when

it became unmanageable because

of the sizes of the artifacts in real life SW developments.

A *hairy* task! (Think “hairy theorem”!)

Needles in Haystack

Like looking for needles in a haystack that is also

a dump for small appliances and electronics.

Tacit Assumption

It was *obvious* to me that

the key measure would be recall.

So obvious that I never put this fact in words.

What the Paper Did

Well… The paper goes on and on about

how important the task is in RE

how hard it is for people to do the task

how important having a good tool is

the design of the tool and

how important it is to empirically evaluate the tool

So far so good!

Gold Set

I read about how

the authors built a gold set

with domain experts working individually on the test data

and then coming to a consensus on the correct answers

that any tool for the task should give.

So far so good!

The Data

I read all about the data.

I see recall of 98% and precision of 50% and I say "Wow!"

Then I am gobsmacked when

I read the authors’ saying

that the tool is not so good because of the low precision

Huh?

Huh?

The tool probably did a whole lot better than I, a human, could have and even though there are some false positives in the output.

False Positives

They are easily weeded out

with a manual search (called vetting)

that is a lot smaller than a manual search of the whole input,

whose recall would not be as high beause of the tedium.

Needles

Think of the needles in a haystack that is also a dump for small appliances and electronics.

A magnet that pulls out all needles and the not too heavy junk

has 100% recall and 50% (say) precision.

I am ecstatic. ☺

Separating Out Needles

Manually separating the needles from

the small appliances and electronics

is a *lot* easier and faster than

manually searching for needles the entire haystack.

Not to mention, less of a pain in the fingers ☺ !!!

Even More Gobsmacky

Sometimes, I see in data, …

recall of 50% and precision of 98%

and I say "Yecchhhhh!" ☹

I am gobsmacked when

I read the authors' saying that

the tool is great because of the high precision.

Huh?

Why It’s Yecchhhh

It failed to find half,

although it did a good job filtering out junk.

But it’s easy to tell junk when I look at it,

but not easy to find the good stuff.

With this tool,

I gotta go and do the whole thing manually anyway.

Think

Not needles in a haystack,

but searching for lost keys under a street lamp,

***not* because that’s where you lost the keys,**

***but* because that’s where the light is ☹ !!**

I Decided

I decided that I had to try to help tool builders with their evaluations.

That work led through several papers and

eventually to the J1st paper in *EMSE*.

☺