

Recall vs. Precision vs. Summarization in RE for AI

Daniel M. Berry
University of Waterloo

Tasks Requiring Intelligence

We are talking about *tasks* requiring *real intelligence (RI)*, i.e., from a *human*.

The task is to find *correct answers* in a space of *answers*, some *correct* and the rest *incorrect*.

Building an AI or LM

We want to build an *artificial intelligence (AI)* that does the task.

This AI might be a *learned machine (LM)* which is the result of *machine learning (ML)*, whether it is taught, self-teaching, or both.

Specifying Requirements of AI

How do we *specify the requirements* of the AI in a way that ...

when we have an *implementaion* of the AI,

we can use the

requirements specification (RS) of the AI

to decide whether the

implementation *meets*

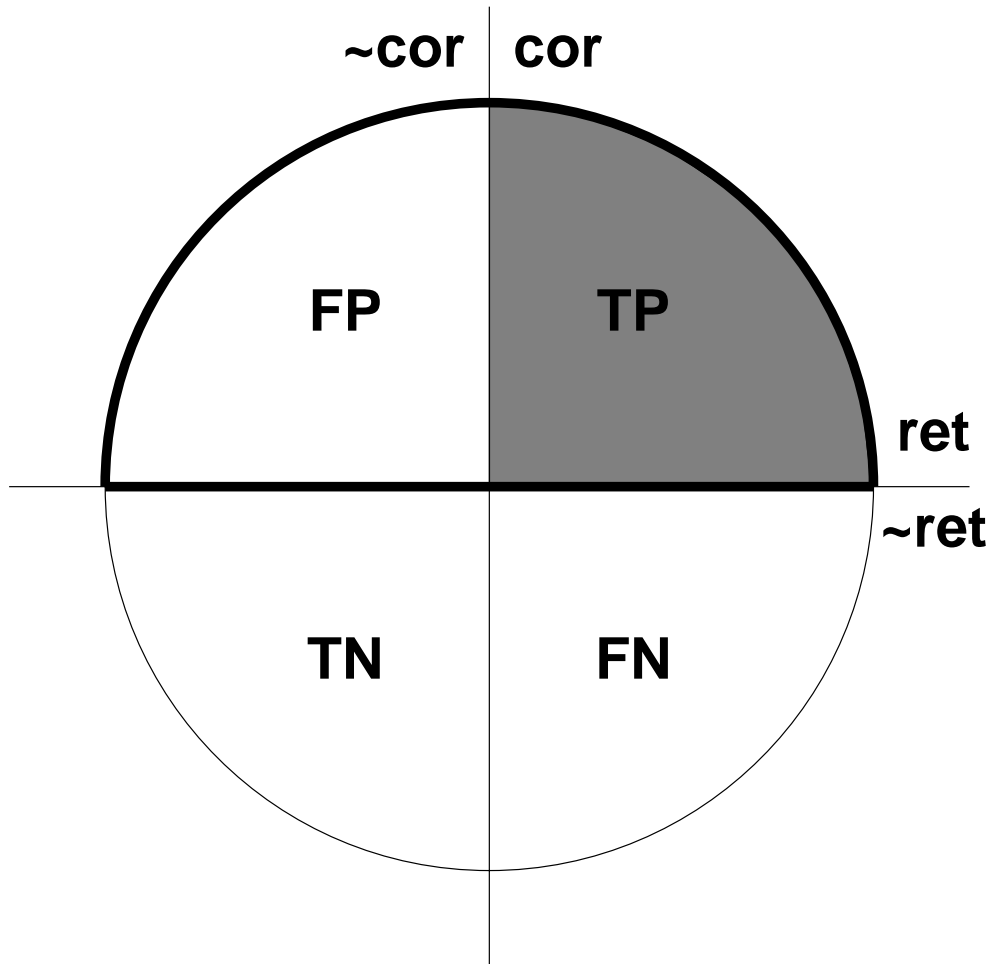
the AI's *requirements*?

Precision

P is the percentage of the tool-returned answers that are correct.

$$\begin{aligned} P &= \frac{| \mathit{ret} \cap \mathit{cor} |}{| \mathit{ret} |} \\ &= \frac{| \mathit{TP} |}{| \mathit{FP} | + | \mathit{TP} |} \end{aligned}$$

Precision

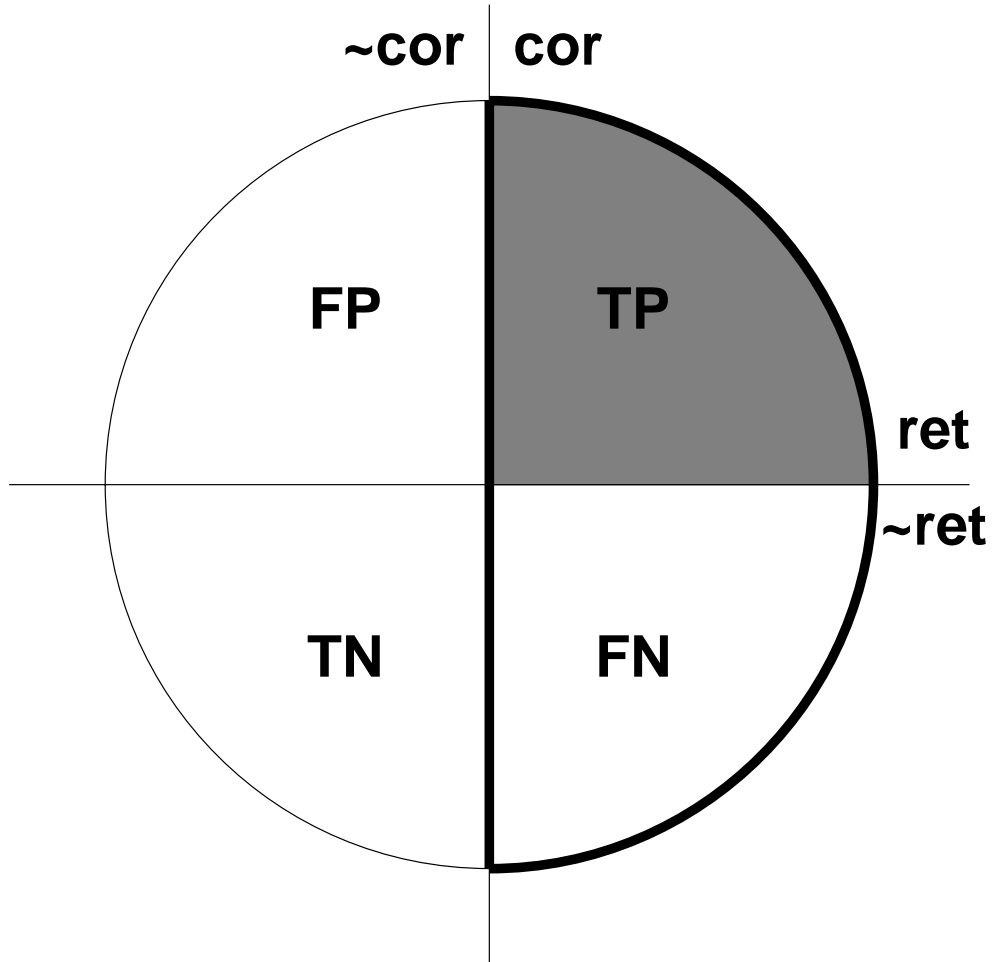


Recall

R is the percentage of the correct answers that the tool returns.

$$\begin{aligned} R &= \frac{| \mathit{ret} \cap \mathit{cor} |}{| \mathit{cor} |} \\ &= \frac{| \mathit{TP} |}{| \mathit{TP} | + | \mathit{FN} |} \end{aligned}$$

Recall



F-Measure

F*-measure: harmonic mean of *P* and *R
(harmonic mean is the reciprocal of the arithmetic mean of the reciprocals)

Popularly used as a composite measure.

$$F = \frac{1}{\frac{\frac{1}{P} + \frac{1}{R}}{2}} = 2 \cdot \frac{P \cdot R}{P + R}$$

Weighted F -Measure

For situations in which R and P are not equally important, there is a weighted version of the F -measure:

$$F_{\beta} = (1 + \beta^2) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R}$$

Here, β is the ratio by which it is desired to weight R more than P .

Note That

$$F = F_1$$

**As β grows, F_β approaches R
(and P becomes irrelevant).**

If Recall Very Very Important

Now, as $\beta \rightarrow \infty$,

$$\begin{aligned} F_{\beta} &\approx \beta^2 \cdot \frac{P \cdot R}{\beta^2 \cdot P} \\ &= \frac{\beta^2 \cdot P \cdot R}{\beta^2 \cdot P} = R \end{aligned}$$

As the weight of R goes up, the F-measure begins to approximate simply R !

If Precision Very Very Important

Then, as $\beta \rightarrow 0$,

$$F_{\beta} \approx 1 \cdot \frac{P \cdot R}{R}$$
$$= P$$

which is what we expect.

R vs *P* Tradeoff

P and *R* can usually be traded off in an IR algorithm:

- increase *R* at the cost of decreasing *P*, or
- increase *P* at the cost of decreasing *R*

Extremes of Tradeoff

Extremes of this tradeoff are:

1. tool returns all possible answers, correct and incorrect: for

$$R = 100\%, P = C,$$

$$\text{where } C = \frac{\# \text{ correctAnswers}}{\# \text{ answers}}$$

2. tool returns only one answer, a correct one: for

$$P = 100\%, R = \varepsilon,$$

$$\text{where } \varepsilon = \frac{1}{\# \text{ correctAnswers}}$$

Extremes are Useless

Extremes are useless, because in either case,

...

the entire task must be done manually on the original document in order to find *exactly* the correct answers.

100% Recall Useless?

Returning everything to get 100% R doesn't save any real work, because we still have to manually search the entire document.

This is why we are wary of claims of 100% R ... Maybe it's a case of this phenomenon!

What is missing?

Summarization

Different Summarization

The summarization I define is different from any semantics-based summarization that you may be thinking of.

I am telling you this now so that you are not surprised when I don't use it in the way you expected.

Summarization

If we can return a subdocument significantly smaller than the original ...

that contains *all* correct answers, ...

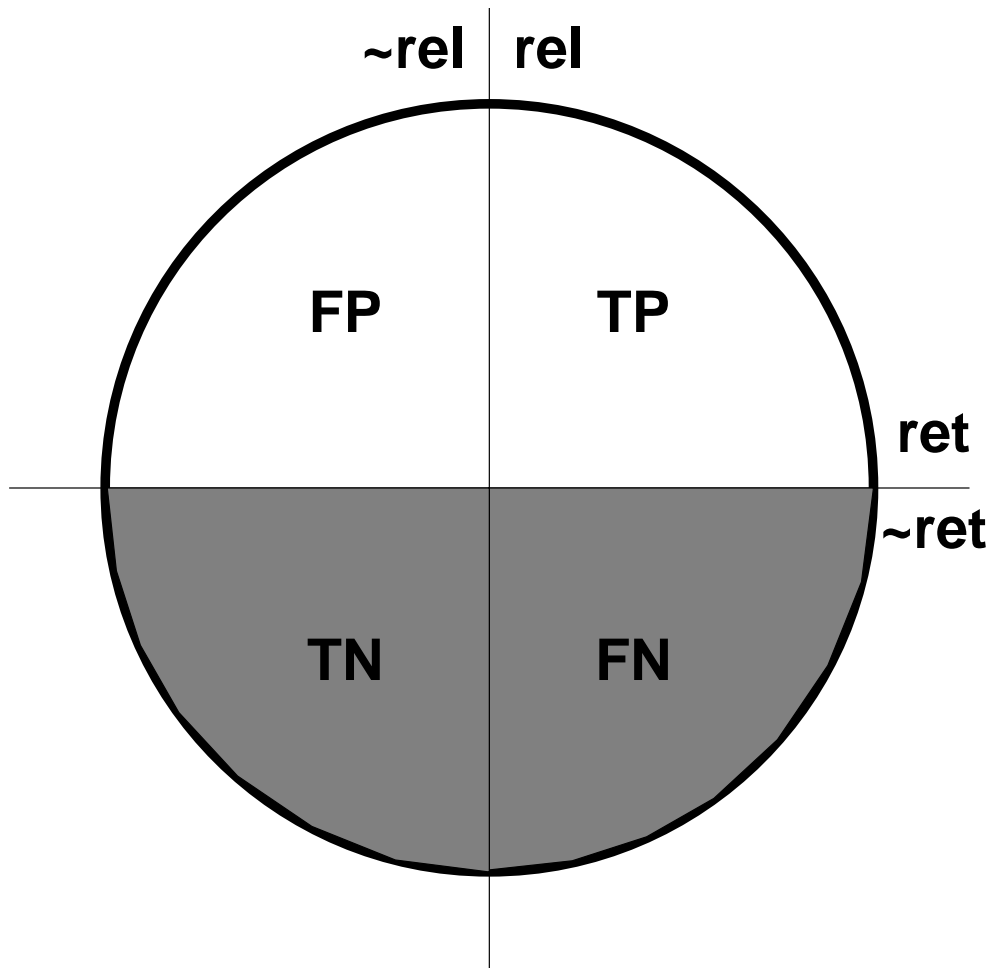
then we have saved some real work.

The *remaining* manual task will take significantly less time than the original, pre-tool-application manual task.

Summarization Measure

Summarization = fraction of the original document that is eliminated in what is returned

$$\begin{aligned} S &= \frac{|\sim ret|}{|\sim ret \cup ret|} = \frac{|\sim ret|}{|\sim rel \cup rel|} \\ &= \frac{|TN| + |FN|}{|TN| + |FN| + |TP| + |FP|} \end{aligned}$$



How to Use Summarization

If there is no escaping doing the task, and the alternative to using a tool is to do the task manually, then ...

we would *love* a tool with 100% *R* and 90% *S*.

Then we really do not care about *P*,

With high *S*, the time to vet the tool's output will be significantly smaller than the time to do the task entirely manually, *and* ...

we end up with potentially the same 100% *R*.

In Other Words

That is, if we can get rid of 90% of the document with the assurance that ... what is gotten rid of contains *only irrelevant* items and thus ...

what is returned contains *all* the relevant items, and

the time to vet the output is only 10% of the time to do the task manually on the input document,

then we are *very happy!* 😊

Remember:

For an AI, A

Recall (R): percentage of the correct stuff that is returned by A

Precision (P): percentage of the stuff returned by A that is correct

Summarization (S): percentage of the input to A that is removed in the output that A returns, i.e., $(100\% - (\text{size}(\text{output}) / \text{size}(\text{input})))$

First Stop-Sign Finder AI

One AI, A1, to classify each image,

Has_a_Stop_sign or not,

and to output only those that have at least one stop sign.

First Stop-Sign Finder AI

A1 picks out of thousands of images the few that contain stop signs in order

to produce a training set of images,

each of which is correctly classified as to whether it contains a stop sign,

to be used to train another AI, A2,

to be used in an autonomous vehicle (AV) to identify stop signs in real time.

Second Stop-Sign Finder AI

Another AI, A2, to classify each of a continuous stream of images,

Has_a_Stop_sign or not,

and if A2 sees a stop sign,

it signals to the AV to stop at the stop sign.

Evaluation of Each AI

Each AI is evaluated by its R and P ,

w.r.t. a manually developed gold set.

Humanly-Achievable R and P

Each human participating in developing the gold set computes his or her own R and P ,

and the average of their R and P values are

**the humanly-achievable-recall (HAR) and
the humanly-achievable-precision (HAP)**

of the stop-sign recognition task.

HAR and HAP

Each of the HAR and HAP of the stop-sign recognition task is probably about 99.99%.

But could probably get very accurate data from www.captcha.net, 😊 .

Basic *A1* Requirement

For *A1*, the basic requirement is to achieve *R* of 100% and *P* of 100%,

because any *R* and *P* less than 100% means

Basic $A1$ Requirement

For $A1$, the basic requirement is to achieve R of 100% and P of 100%,

because any R and P less than 100% means

that the data on which $A2$ trains are flawed,

and $A2$ will not learn perfectly,

setting up possible AV failures.

Basic *A2* Requirement

For *A2*, the basic requirement is to achieve *R* of 100% and *P* of 100%,

because any *R* less than 100%, e.g., *R'* means

Basic A2 Requirement

For A2, the basic requirement is to achieve R of 100% and P of 100%,

because any R less than 100%, e.g., R' means

that the AV will fail to stop at 100% – R' of the stop signs.

Basic *A2* Requirement

For *A2*, the basic requirement is to achieve *R* of 100% and *P* of 100%,

because any *P* less than 100%, e.g., *P'* means

Basic *A2* Requirement

For *A2*, the basic requirement is to achieve *R* of 100% and *P* of 100%,

because any *P* less than 100%, e.g., *P'* means

that the *AV* will stop unnecessarily 100% – *P'* of the time.

100% R and P Achievable?

But, are $R = 100\%$ and $P = 100\%$ achievable?

100% R and P Achievable?

But, are $R = 100\%$ and $P = 100\%$ achievable?

No

Are $R = 100\%$ and $P = 100\%$ even reasonable to expect?

100% R and P Achievable?

But, are $R = 100\%$ and $P = 100\%$ achievable?

No

Are $R = 100\%$ and $P = 100\%$ even reasonable to expect?

No

What *Is* Achievable?

So what *are* reasonable *R* and *P* to expect to achieve and maybe beat?

What *Is* Achievable?

So what *are* reasonable *R* and *P* to expect to achieve and maybe beat?

HAR and HAP (what is humanly achievable)

What *Is* Achievable?

So what *are* reasonable *R* and *P* to expect to achieve and maybe beat?

HAR and HAP (what is humanly achievable)

For *A1*, the basic requirement is to meet or exceed *R* of HAR and *P* of HAP.

For *A2*, the basic requirement is to meet or exceed *R* of HAR and *P* of HAP.

Vetting of an AI's Output?

The output of *A1* is vetted (checked) by human beings.

There is plenty of time, and vetting improves overall *R* and *P*.

The output of *A2* is *not* vetted by human beings.

There is no time, because the output is needed *immediately* by its AV.

Amended Basic Requirement

For *each* of $A1$ and $A2$,

R should be at least as good as HAR, and

P should be at least as good as HAP,

Can we do better with both R and P ?

Tradeoff

Unlikely!

In implementing any AI, a higher R can be achieved at the cost of lowering P , and vice versa:

Tradeoff

A less strict classification causes more images to be accepted as containing a stop sign, thus

increasing R , and unfortunately,

increasing imprecision (= $100\% - P$).

Tradeoff

A more strict classification causes fewer images to be accepted as containing a stop sign, thus

**increasing P , and unfortunately,
decreasing R .**

Usually Recall is Critical

Regardless of how critical a high P is,
a high R is very critical.

Finding *all* correct answers is often *very*
necessary.

Lives depend on doing so.

If finding *all* correct answers were not
necessary, we would not bother building an AI
for doing it.

If High R is Important

However, if achieving high R is important, there may be *no* choice but to accept low P .

For $A1$, what does low P mean?

If High R is Important

However, if achieving high R is important, there may be *no* choice but to accept low P .

For $A1$, what does low P mean?

Lots of false positives among the output of $A1$

Are these false positives dangerous?

If High R is Important

However, if achieving high R is important, there may be *no* choice but to accept low P .

For $A1$, what does low P mean?

Lots of false positives among the output of $A1$

Are these false positives dangerous?

No

Effect of False +s on Veters

What effect do lots of false positives among the output of A1 have on veters?

Very discouraging

Effect of False +s on Vectors

How tolerable is low P for $A1$?

Effect of False +s on Vectors

How tolerable is low P for $A1$?

if S is high?

Effect of False +s on Vectors

How tolerable is low P for $A1$?

if S is high?

very tolerable

if S is low?

Effect of False +s on Vectors

How tolerable is low P for $A1$?

if S is high?

very tolerable

if S is low?

not tolerable

Low P for $A1$

In the end, for $A1$,

**if the effective R of $A1$ determined after vetting
beats the task's HAR,**

**and the time to vet $A1$'s output is less than the
time to do the classification task manually,**

in other words, to effectively vet $A1$'s input,

then $A1$ meets its requirements.

Low P for $A1$

**After all, since the task of $A1$ is essential,
the alternative to running $A1$ is
to do the task completely manually...**

**at the cost of a lot *more* tedious, *boring* grunt
work!**

yechhhhh!!!!

Effect of False +s on AVs

For A2, what does low *P* mean?

Remember that there is *no* vetting.

A2 runs in an AV.

Effect of False +s on AVs

For A2, what does low P mean?

Remember that there is *no* vetting.
A2 runs in an AV.

Lots of unnecessary stops by the AV!

Are these unnecessary stops by the AV dangerous?

Effect of False +s on AVs

For A2, what does low P mean?

Remember that there is *no* vetting.
A2 runs in an AV.

Lots of unnecessary stops by the AV!

Are these unnecessary stops by the AV
dangerous?

Could very well be; could lead to lots of rear-
end-from-front-end collisions!

Low P for $A2$

How tolerable is low P for $A2$?

Low P for $A2$

How tolerable is low P for $A2$?

Definitely not!

Low P for $A2$

How tolerable is low P for $A2$?

Definitely not!

In the end, for $A2$,

if R of $A2$ meets or beats the HAR,
and P of $A2$ meets or beats the HAP,
then $A2$ meets its requirements.

Ease of Implementation

So,

implemenation of A1

will be significantly easier than

implementation of A2.

Conclusion

It's clear that a requirements specification (RS) for an AI needs more than just *R* and *P*.

Needs More Than R And P

RS needs also:

- **HAR of the task, with which to compare R ,**
- **HAP of the task, with which to compare P ,**
- **β of the context of the AI, the ratio by which to weight R more than P , and**
- **in a case in which vetting is possible or required, S of the AI, to help evaluate the tradeoff between R and P .**

Requires Understanding Context

The second last value requires a *full* understanding of the context in which the AI is being used, including

- the cost of a false negative, and
- the cost of a false positive

in the context.

Decision Not Cut And Dry

**Finally, the decision of whether
the AI satisfies its RS and meets its
requirements
will involve
engineering judgement and
evaluation of tradeoffs in the AI's context,
and will *not* be a simple “yes” vs “no”
decision.**

Decision Not Cut And Dry

The decision is not cut and dry because of all of the vague elements in the RS.

The RS for an AI is as vague as are fitness criteria for vague qualitative, non-functional requirements, e.g., “fast response time” or “friendly user interact”.

Vagueness in the RS

For examples:

1. How much *S* is enough for vetters to *tolerate* having to vet in the presence of *low P*?
2. How *critical* must the task be in order that the *only* alternative to an AI that meets its RS is doing the task manually?
3. How much *S* is enough for vetters to *tolerate* having to vet *more than* having to do the task manually?

Vagueness in the RS

What is the interaction between the previous three questions?

What is to be done in the situation in which the task is *fairly critical*, the AI *just misses achieving* the task's HAR and HAP, but the S is *very high*?

Continually Learning LM

Consider evaluating R and P using a gold set for a LM, M , that learns from its mistakes.

What is the main problem with evaluating every new version of M with the same gold set?

Continually Learning LM

Consider evaluating R and P using a gold set for a LM, M , that learns from its mistakes.

What is the main problem with evaluating every new version of M with the same gold set?

M learns from its mistakes on the gold set, and the gold set loses validity as an instrument.

Continually Learning LM

OK, So we use a different gold set for each new version of M !

What is wrong with doing so?

Continually Learning LM

OK, So we use a different gold set for each new version of M !

What is wrong with doing so?

We lose comparability of the evaluations that comes from knowing that all versions are evaluated against the same gold set,

and now have to trust that all gold sets are equally difficult to any LM.

Continually Learning LM

**So what do we have to do to preserve validity
and comparability?**

Continually Learning LM

**So what do we have to do to preserve validity
and comparability?**

**Somehow inhibit M 's learning from its
mistakes on the gold set.**

Acknowledgments

Thanks to Jo Atlee for her comments on an earlier version of these slides.