Requirements for Requirements Engineering Tools that Require Understanding Requirement Semantics

...

Why such tools should be clerical and not NLP-based

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

Requirements for Tools for Hairy Requirements or Software Engineering Tasks

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

Vocabulary

CBS = Computer-Based System

SE = Software Engineering
RE = Requirements Engineering
RS = Requirements Specification

NL = Natural Language
NLP = Natural Language Processing
IR = Information Retrieval

HD = High Dependability

HT = Hairy Task

Hairy Task (HT)

A hairy RE or SE task involving NL documents:

requires NL understanding *and*
is not difficult for humans to do on a small scale *but*
is unmanageable when it is done to the
documents or artifacts that
accompany the development of a large
CBS.

Examples of HTs

Examples include finding

- **abstractions,**
- **ambiguities, and**
- **trace links**

I chose the word “hairy” to evoke the metaphor of the hairy theorem or proof.

HTs Need Tool Support

A hairy task (HT) is burdensome enough that humans need tool assistance to do complete job.

Humans understand NL well enough that
a human has the potential of
achieving for the HT task
close to 100% *correctness*,

i.e., of finding close to all and only the desired information.

Correctness

Two components of "correctness" are

- ***recall*, that all the desired information is found, and**

- ***precision*, that only the desired information is found.**

Recall vs. Precision

Of recall and precision, for a HT, recall is more in need of tool assistance.

Finding a unit of desired information among the many documents and artifacts available
for the CBS's development
is generally significantly harder than dismissing a found unit of information
that is not desired.

Therefore, …

Therefore, for a HT,
if close to 100% correctness is needed,
then close to 100% recall is needed.

Perfection Not Always Needed

Not every instance of a HT for the development of a CBS needs to achieve close to 100% recall.

However, if the CBS being developed has HD requirements, then recall for the HT must be as close as possible to 100% in order to ensure that the HD will be achieved [BGST12].

HD Case

E.g., 100% of all trace links must be found in order to *ensure* that all the effects of any proposed change can be traced.

If Not

In this HD case,
if a tool for the HT achieves less than
close to 100% recall,
then the task must be done manually on all
the docs to find
the links that the tool does not find.

Therefore, in the last analysis, such a tool is really useless.

Maybe Not Totally Useless

Could argue that even such a tool is useful as a defense against a human’s <100% recall, using the tool as a double check after the human has done the tool’s task manually.

But, I believe that if the human *knows* that the HT tool will be run, the human might be lazy and not do the HT manually as well as possible.

Empirical Studies Needed

Empirical studies are needed to see
if this effect is real, and
if so, how destructive it is
of the human's recall.

How Close to 100% Recall?

Just how close to 100% must the recall of a tool for a HT be?

We know that

1. **a human’s achieving 100% recall is probably impossible**

We know that, Cont’d

2. even if achieving 100% recall were possible,

   there is no way to know if we have
   succeeded,
   because the only way to measure recall
   for a tool
   is to compare the output of the tool against
   totally correct output,
   which can be made only by humans.

Actual Human Recall

Consider a human performing a HT manually under the best of conditions.

Let's call the best recall that the human can achieve the "humanly achievable high recall (HAHR)", which we hope is close to 100%.

a.k.a. "the gold standard for evaluating tools in NLP

Real Recall Goal for HT

So our real goal for a tool for a HT:

to show that the tool for the HT
measurably achieves
better recall than the HAHR for the HT.

So there is some empirical work to be done, at the very least to measure for each HT its HAHR.

Acceptable Recall for HT Tools

What about tools for HTs?

If a tool for a HT gets better recall than HAHR, then a human

will trust the tool and

will not feel compelled to do the HT manually

to look for what the tool missed.

So there is more empirical work to be done, to measure each tool’s recall.

Not All Tools Work Alone

In general, a tool may work best or may be designed to work with humans.

If so, the recall of the tool is not the raw recall of the tool, but the recall of a human working with the tool.

Evaluate a Tool with Human

In general, a tool for a HT must be evaluated by comparing

the recall of humans working with the tool with

the recall of humans carrying out

HT manually.

Empirical Evaluation

Therefore, the evaluation of any tool for a HT

requires an experiment comparing

- **application of the tool to the HT,**
  - **with or without human help**
- **with**
  - **humans' doing HT completely**
    - **manually.**

Natural Language in RE

Getting back to NLs in RE, …

A large majority of requirements specifications (RSs) are written in natural language (NL).

Tools to Help with NL in RE

For nearly 30 years, there has been much interest in developing tools to help analysts overcome the shortcomings of NL for producing precise, concise, and unambiguous RSs.

Many of these tools draw on research results in NL processing (NLP) and information retrieval (IR) (which we lump together under “NLP”).

NLP-Based Tools and RE

NLP research has yielded excellent results, including search engines!

This talk argues that characteristics of RE and some of its tasks impose requirements on NLP-based tools for them and force us to question whether …

for any particular RE task, is an NLP-based tool appropriate for the task?

Categories of NL RE Tools

Most NL RE tools fall into one of 4 broad categories (a–d):

a. **finding defects and ambiguities in NL RSs,**
b. **generating models from NL descriptions,**
c. **finding trace links among NL artifacts and other artifacts,**
d. **finding key abstractions in NL pre-RS documents,**

Three of these, a, c, and d, are HTs!

Key Needed Capability of Tools

Except for an occasional tool of category (a), part of whose task may include format and syntax checking …

each RE task supported by the tools requires *understanding* the contents of the analyzed documents.

Can Tools Deliver Capability?

However, understanding NL text is still way beyond computational capabilities.

Only a very limited form of semantic-level processing is possible [Ryan1993].

“I Know I’ve Been Fakin’ It”

Consequently, most NLP RE tools …

use mature techniques for identifying lexical or syntactic properties, and …

then *infer* semantic properties from these.

That is, they *fake* understanding.

Limitations of NLP-Based Tools

Limitations of NLP-Based Tools for HTs

Typical tool for a HT is built using NL processing (NLP),

involving at least a parser and a parts-of-speech tagger (POST)

Limitations, Cont'd

Even the best parsers are no more than 85–91% accurate [SBMN13].

Even the best parts-of-speech tagger are no more than 97.3% accurate [Manning11].

No NLP-based tool can be better than the worse of its parser and its tagger.

No NLP-based tool will achieve more than 85–91% recall.

Fundamental Limitation

This is the fundamental limitation of NLP-based tools for HT, which is problematic because:

NL text that is found in real-life software development documents is sloppy and is inherently ambiguous and anomalous.

New Approaches for Tools

If we have time at the end, we will examine several alternative approaches for building tools for HTs.

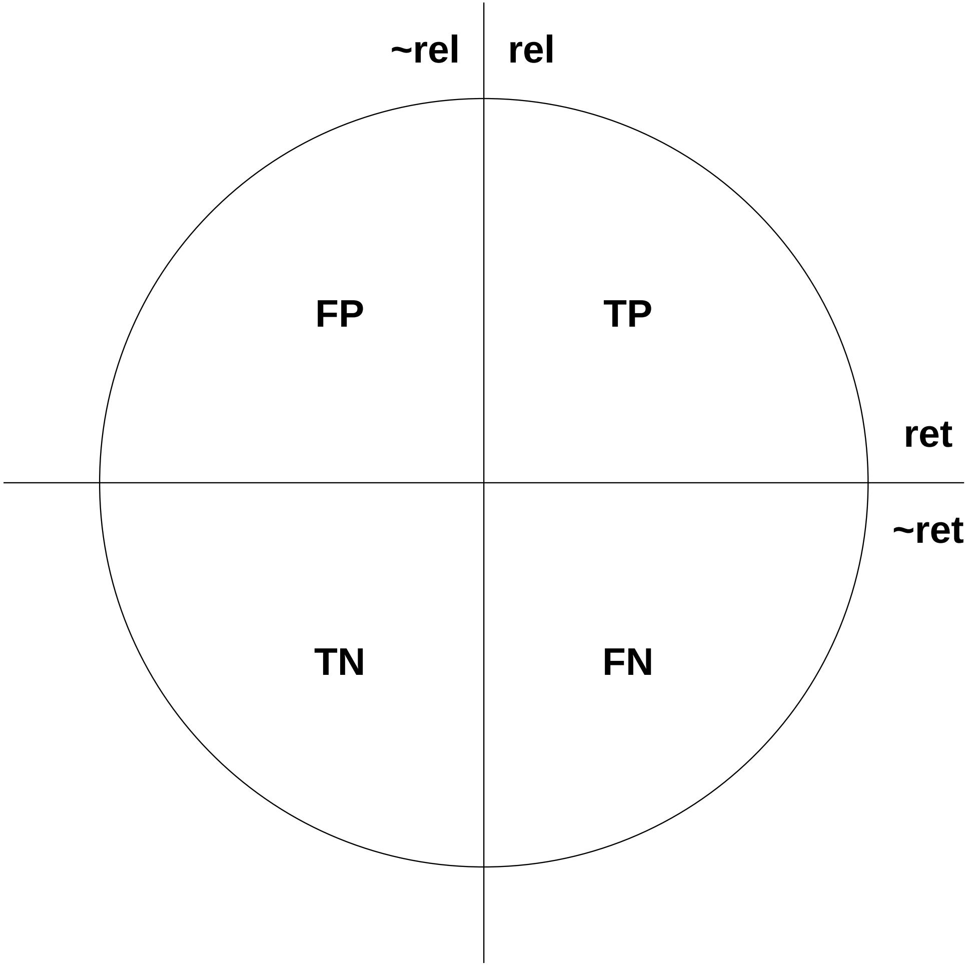
New Approaches, Cont’d

For now, I will only mention only two:

- **Algorithmic partitioning of the HT into clerical and hairy parts,**
    - **building a tool with 100% recall for the clerical part and**
    - **letting humans do hairy part manually, ignoring the clerical part, but possibly using the tool’s output.**
- **Machine learning (We are seeing recently that ML can achieve close to HAHR.)**

Measures to Evaluate Tools

The Universe of an RE Tool

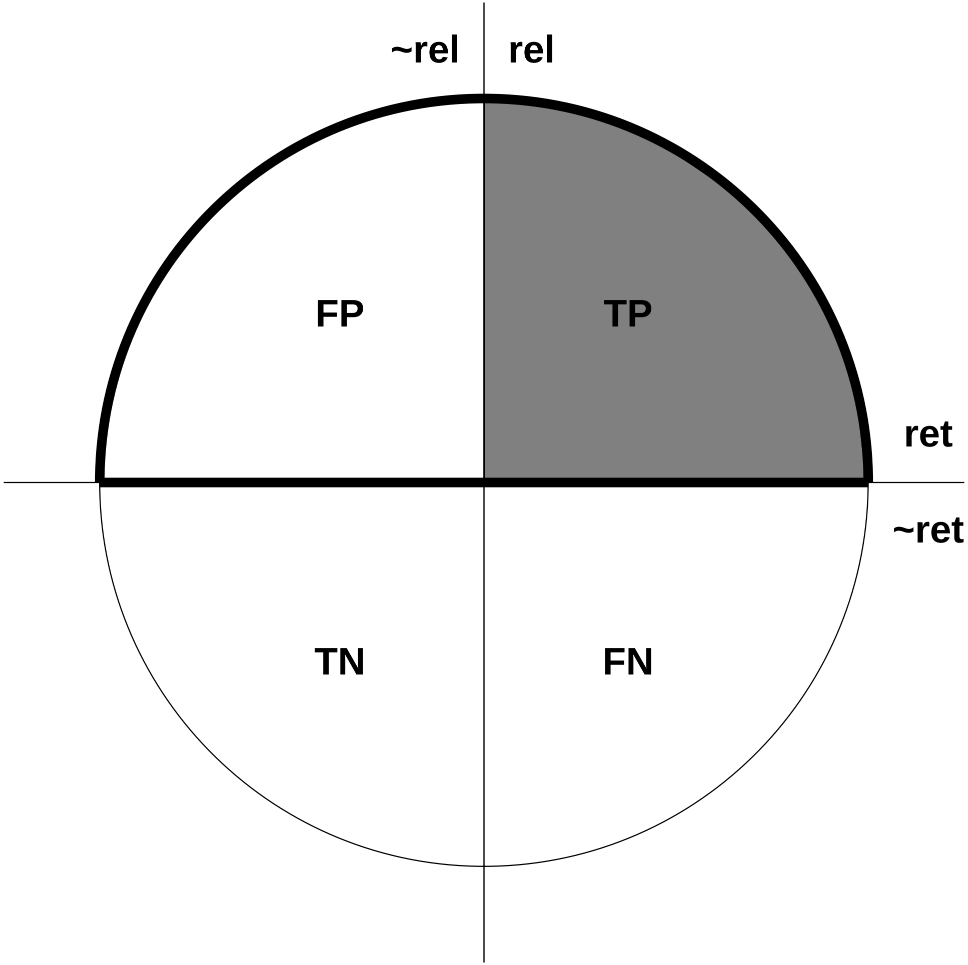


Precision

Precision: fraction of the retrieved items that are relevant

$$P = \frac{|ret \cap rel|}{|ret|}$$

$$= \frac{|TP|}{|FP| + |TP|}$$

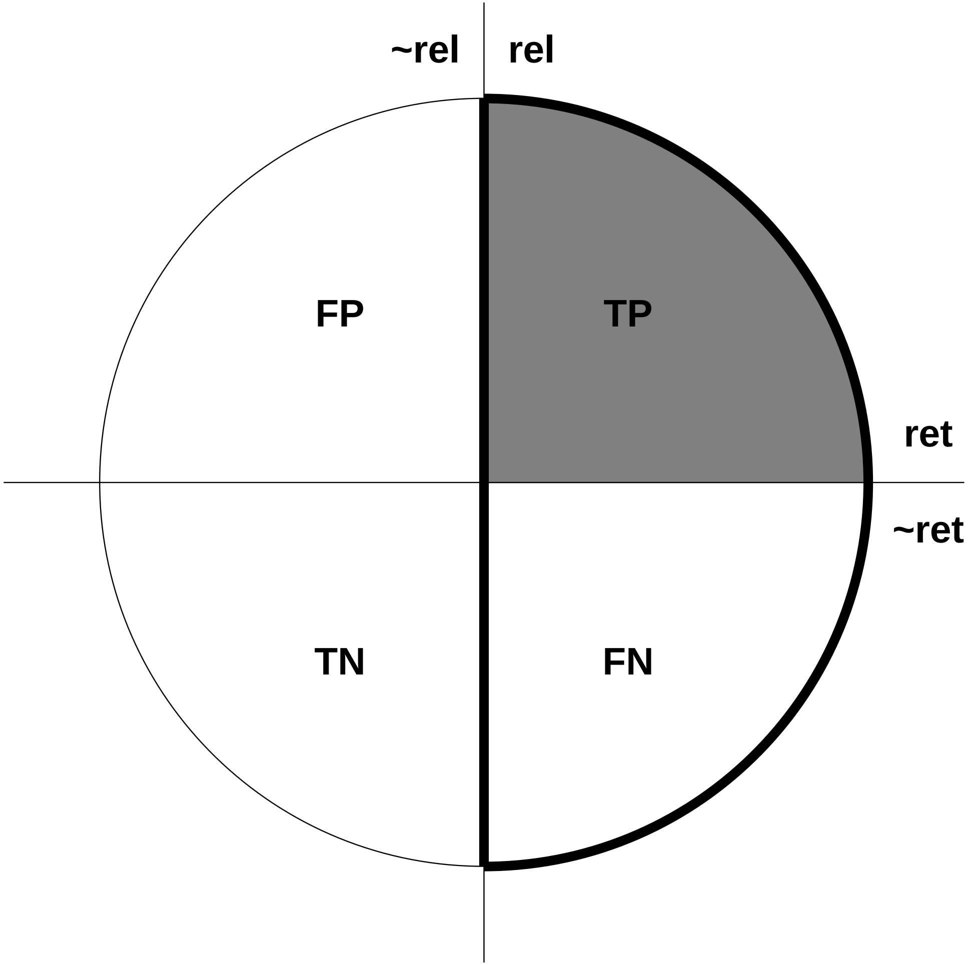

~rel
rel
FP
TP
ret
~ret
TN
FN

Recall

Recall: fraction of the relevant items that are retrieved

$$R = \frac{|ret \cap rel|}{|rel|}$$

$$= \frac{|TP|}{|TP| + |FN|}$$


~rel
rel
FP
TP
ret
~ret
TN
FN

F-Measure

F-measure: harmonic mean of precision and recall (harmonic mean is the reciprocal of the arithmetic mean of the reciprocals)

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

Popularly used as a composite measure

Incorrect Assumption

But this assumes that *P* and *R* carry the same weight.

However, for a typical HT, manually finding a missing correct answer (a false negative)

is significantly harder than

rejecting as nonsense an incorrect answer (a false positive), …

Reality, Because

because finding a missing correct answer generally requires examining all the input documents in detail,

while

rejecting an incorrect answer generally requires understanding only the incorrect answer and the input documents at only a general level [KHDH11]

Footnote: Essential Hairiness

If fact, it seems reasonable to include in the definition of a HT.

the proviso that manually finding a true positive or false negative is

> **significantly harder than**
> **rejecting a false positive.**

Any task for which this difficulty difference is not true does not satisfy the unmanageability criterion of the definition.

Recall vs. Precision?

In summary, …

for a tool for a HT,

recall appears to be at least an order of magnitude more important than precision, …

especially when the tool is applied to the artifacts of a HD CBS.

Weighted Harmonic Mean

So let’s do a weighted mean harmonically, with *w* as the weight of *R* over *P*

$$F_W = \frac{1}{\dfrac{\dfrac{1}{P} + w \cdot \dfrac{1}{R}}{w+1}}$$

$$F_W = (w+1) \cdot \frac{P \cdot R}{w \cdot P + R}$$

Note that $F = F_1$.

Recall = 10 × Precision

To reflect that recall is at least an order of magnitude more important than precision, let $w = 10$.

$$F_{10} = 11 \cdot \frac{P \cdot R}{10 \cdot P + R}$$

Note that $F_{\frac{1}{10}}$ weights P ten times over R

I Do Not Understand

I do not understand why the literature on the F-Measure uses the square in the weighted formula

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

to weight R β times P.

How should β be determined?

It should be calculated as some function of

1. **an estimate of the ratio of the time for a human to manually find a true positive in the original documents and the time for a human to reject a tool-presented false positive, and**
2. **an estimate of ratio of the cost of the failure to find a true positive and the cost of the accumulated nuisance of dealing with tool-presented false positives.**

Determining β, Cont'd

For any particular HT, a separate empirical study is necessary to arrive at good estimates for these ratios.

If Recall Very Very Important

Now, as $w \rightarrow \infty$,

$$F_W \approx w \cdot \frac{P \cdot R}{w \cdot P}$$

$$= \frac{w \cdot P \cdot R}{w \cdot P} = R$$

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

If Precision Very Very Important

Then, as $w \to 0$,

$$F_W \approx 1 \cdot \frac{P \cdot R}{R}$$

$$= P$$

which is what we expect.

Recall vs. Precision

Many a tool for a HT is reported *happily* in the literature as having more precision than recall [DRS13].

**Sometimes, a tool that has precision = 85%
and that has recall = 65%
is reported as satisfactory [GZ14].**

Huh?!?!

Why Do We Love Precision?

Why is there such an emphasis on precision?

Precision is important in the information retrieval area from which are borrowed many of the algorithms used to construct the tools for HTs.

In information retrieval, users of a tool with low precision are turned off by having to reject false positives more often than they accept true positives.

Why Do We Love …, Cont'd?

In some cases, only a few or even only one true positive is needed.

Perhaps the force of habit drives people to evaluate the tools for HTs with the same criteria that are used for information retrieval tools.

Also, "precision" sounds so much more important than "recall", as in "This output is precisely right!".

Tradeoff

For a typical RE task in which finding relevant items is at least an order of magnitude harder than rejecting irrelevant items, it pays to sacrifice precision for recall.

But …

The Extreme Tradeoff

Return …

the entire document → R = 100% & P = 0%

nothing → P = 100% & R = 0%

Useless

But returning everything to get 100% recall doesn’t save any real work, because we still have to manually search the entire document.

What is missing?

Summarization

Summarization

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

that contains *all* relevant items, …

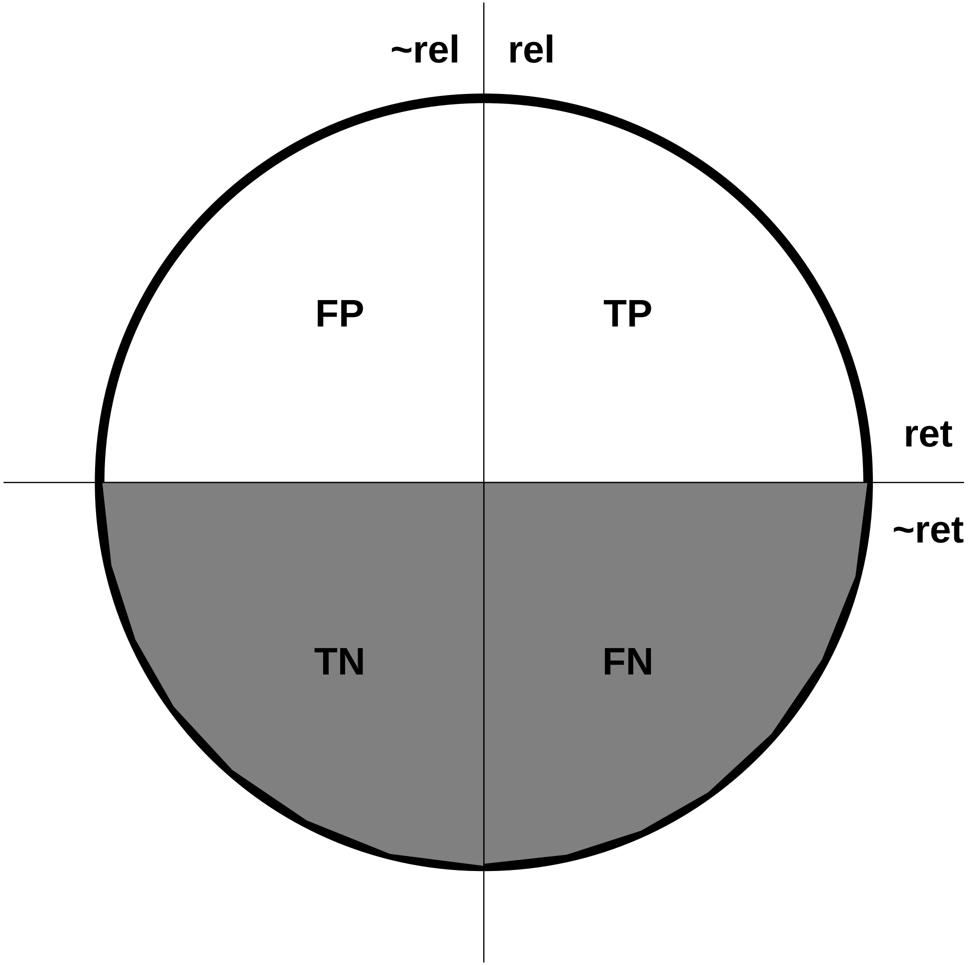
then we have saved some real work.

Summarization Measure

Summarization = fraction of the original document that is eliminated from the return

$$S = \frac{|\sim ret|}{|\sim ret \cup ret|} = \frac{|\sim ret|}{|\sim rel \cup rel|}$$

$$= \frac{|TN| + |FN|}{|TN| + |FN| + |TP| + |FP|}$$


~rel
rel
FP
TP
ret
~ret
TN
FN

How to Use Summarization

We would *love* a tool with 100% recall and 90% summarization.

Then we really do not care about precision.

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, …

then we are *very happy*! ☺

Digression

We now look at some published studies that weight precision and recall equally, …

but whose results can be improved by weighting recall at least 10 times precision.

Conclusion

Most RE tasks involving NL documents are HTs.

Tool support for them is essential, *because* of the hairiness.

The hairiness of these tasks makes high recall essential.

We have built mostly NLP-based or IR-based tools for these HTs.

Conclusion, Cont'd

But, the HTs' very hairiness makes tools for them have less recall than humans are capable of on a small scale.

From force of habit in NLP and IR fields, we have been evaluating these tools incorrectly, weighting precision far more than it should be against recall.

This habit has to stop!

New Approach Needed for Tools

Since an NLP-based tool cannot achieve better than 85–91% recall,

perhaps it is time to try other approaches to design a tool for a HT.

An examination of the RE and SE tools literature shows a number of promising approaches worth pursuing.