

Report 6: Rules Extractor General Improvements

Aliyan Faisal · general changes to how the rules extractor reads text, scored on the 30 invoices of Report 5

These changes were made while looking at the 30 invoices they are scored on. The new score shows that the problems found in Report 5 are gone. It is not a fresh measure; a fair figure needs a new set of invoices that played no part in the changes.

1. Summary

Report 5 showed that the rules extractor fails on layouts nobody wrote a rule for: prose lists, e-mail text and cards. For this report I did not write a rule for each layout. I made a small number of general changes to how the extractor reads text, and scored the same 30 invoices in the same way, so the numbers can be compared directly.

Rules extractor, corpus D (30 invoices)	Report 5 (before)	Report 6 (after)
Field accuracy	62.0%	98.9%
Planted mistakes found	15 of 26	26 of 26
False alarms	30	8
Correct invoices with no findings	5 of 15	15 of 15
Fields read wrong or missing	423 of 1098	12 of 1098

Result. Field accuracy rose from 62.0% to 98.9%, all 26 planted mistakes are found, and false alarms fell from 30 to 8. Every line-item cell is now read correctly (177 of 177 in each of the four line columns).

2. What changed

None of these changes names a layout. Each one is a way of reading text that holds for any invoice.

Problem found in Report 5	General change	Effect
Several cells typed into one line of text (a prose list, e-mail text)	A line of text is cut into cells where there is a wide gap, at "quantity × price = amount", and between a label and an amount. Dates and tax IDs are left whole.	Prose 12% to 100%, e-mail 16% to 97%
A line item drawn over two rows (description and amount on one, quantity and price under it)	Two rows are one line when their numbers multiply out.	Cards 30% to 100%
Shipping guessed from a leftover amount	A leftover amount is not taken as shipping when the other figures already add up to the total.	Spurious shipping amounts fell from 6 to 1
Quantity and unit price swapped when the amount column comes first	When quantity × price gives the amount either way round, the quantity is the column of smaller numbers.	Amount-first layout 65% to 98%
Totals whose caption is printed after the amount, or beside it with no table	The caption is taken from either side of the amount.	Totals found on all 30
Supplier read from a title or a column heading	A name ending like a company ("Ltd", "S.r.l.", "Lda") is preferred, a "From:" label is used when there is one, and lines starting with the word for invoice are skipped. A customer line is cut before the address.	Supplier 16 of 30 to 30 of 30
Invoice number, customer and dates not found	Words for these labels in more languages (Portuguese, Spanish, Italian, Dutch, Swedish), and a value may sit above its label as well as beside or below it.	Invoice number 16 to 28 of 30
Tax IDs under local names, and amounts with a country letter	More tax-ID names (for example CNPJ, TRN, RFC), and symbols such as S\$, A\$, C\$ and R\$ are read as money and give the currency.	Tax ID 23 to 30 of 30

3. Results on corpus D

Field	Report 5	Report 6
Line quantity	91/177	177/177
Line unit price	91/177	177/177
Supplier name	16/30	30/30
Invoice number	16/30	28/30
Customer name	17/30	29/30
Tax rate	17/30	30/30
Subtotal	18/30	30/30
Line description	109/177	177/177
Line amount	109/177	177/177
Tax amount	19/30	30/30
Total	19/30	30/30
Tax ID	23/30	30/30
Shipping	24/30	29/30
Currency	26/30	28/30
Discount	26/30	30/30
Issue date	27/30	27/30
Due date	27/30	27/30

Page structure	Report 5	Report 6
prose list	12.4%	100.0%
e-mail	15.5%	96.8%
cards	30.3%	100.0%
receipt	60.7%	99.5%
amount first	65.2%	97.5%
tax column	86.7%	98.8%
summary first	88.9%	98.1%
grouped	89.7%	99.0%
quantity first	94.5%	100.0%
sidebar	94.8%	99.4%

4. The older sets, scored again

The same extractor was run on the three older sets to check that the changes did not make them worse.

Set	Field accuracy (Report 4 / now)	Mistakes found (Report 4 / now)	False alarms (Report 4 / now)
Corpus A: 30 independent invoices	99.07% / 99.01%	31 of 32 / 31 of 32	4 / 4
Corpus B: 40 generator invoices	99.91% / 99.83%	20 of 20 / 20 of 20	0 / 0
Corpus C: 40 multi-page invoices	99.98% / 99.98%	20 of 20 / 20 of 20	0 / 0

The older sets stayed within a tenth of a percentage point of where they were, and no planted mistake that was found before is lost.

5. What is still wrong

12 fields out of 1,098 are still wrong or missing on corpus D, on 7 of the 30 invoices. They are not chased here, because each fix would be written for the invoice in front of me:

Remaining problem	Why it stays
A bare "\$" on a Mexican-peso or Singapore-dollar invoice read as US dollars	Nothing printed says which dollar it is.
Day and month swapped when every date on the invoice has both numbers 12 or below	The order cannot be told from the numbers alone.
Dates written out in Portuguese ("14 de agosto de 2026")	Month names are not read in every language yet.
An invoice number printed with a leading "#" in two layouts, and a customer name cut short	Small gaps left in the reading of those labels.
One spurious shipping amount in a layout with a tax column	The leftover amount fits the arithmetic in a way that looks like shipping.

Conclusion

A handful of general changes took the rules extractor from 62.0% to 98.9% on the 30 invoices of Report 5, with every planted mistake found, and left the older sets where they were. The gap between the rules extractor and the vision model in Report 5 was mostly a gap in how text is split into cells, and that part is now closed for these ten kinds of document.

Read the number with care. I made the changes while looking at these 30 invoices, so the score is higher than it would be on invoices the extractor has not seen. The invoices are synthetic and were written by the same person who built the tool. The next step is a new set of invoices in layouts and languages that played no part in these changes, scored once.

The raw results for this report can be downloaded from the Harness Reports page of this site.