

Report 8: Rules Extractor Improvements on Corpus E

Aliyan Faisal · general improvements to the rules extractor after Report 7, with ten invoices kept aside

Most of these invoices were used to make the fixes. I fixed the extractor using invoices 1 to 40 of corpus E and kept invoices 41 to 50 aside, one for each of my ten layouts, and did not look at their results while fixing. That gives a score that was not tuned on. It is small (ten invoices), the ten layouts and ten countries also appear among the 40 used for the fixes, and I had seen which kinds of fields failed before I split the set. It shows that the fixes work on other invoices of the same kinds, not on layouts or languages nobody has seen.

1. Summary

Report 7 scored the rules extractor on 50 new invoices and found its weak points: currency signs it did not know, label words in other languages, and a few structures. I fixed those in general ways and scored the same 50 invoices again.

Rules extractor	Report 7 (before)	Report 8 (after)
All 50 invoices: field accuracy	92.2%	99.9%
All 50: planted mistakes found	24 of 37	37 of 37
All 50: false alarms	53	2
Invoices 1 to 40 (used for the fixes): field accuracy	94.7%	99.94%
Invoices 41 to 50 (kept aside): field accuracy	64.9%	100.0%
Invoices 41 to 50: planted mistakes found	2 of 6	6 of 6
Invoices 41 to 50: false alarms	15	0

Result. Field accuracy on corpus E rose from 92.2% to 99.9%, all 37 planted mistakes are found, and false alarms fell from 53 to 2. On the ten invoices kept aside the score rose from 64.9% to 100.0%. Every line-item cell is read correctly (813 of 813 in each of the four line columns).

2. What changed

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

Problem found in Report 7	General change	Effect
Amounts followed or preceded by a currency sign it did not know (zl, Kc, TL, kr., yen, rupee)	These signs are read as part of an amount, and give the currency. A bare "kr" is told apart as Danish, Norwegian or Swedish from the tax and language words on the page. Indian digit grouping (1,23,456.00) is read.	Totals, tax, subtotal and currency found on every invoice
Label words in other languages	Words for tax ID, customer, invoice number, due and issue date, subtotal, tax, discount, shipping and total in Polish, Czech, Turkish, Finnish, Danish, Norwegian and Chinese added. A colon left in front of a tax ID is removed.	Tax ID 9 to 30 of 30 on the new-layout invoices; customer, invoice number and dates likewise
Dates written out in more languages	Many more month names, the "de" in "14 de agosto de 2026", and Chinese dates (2026年9月21日).	Issue and due dates found on every invoice; also fixes the Portuguese dates left over from corpus D
Rows with dots between the description and the amount	A row such as "Desk (7 × 1 232,85) 8 629,95" is cut into description, quantity, price and amount.	Dot-leader invoices 11% to 100%
Totals printed before the items	When no totals follow the items, the rows above them are searched.	Totals-before-items 73% to 100%

Problem found in Report 7	General change	Effect
Tax printed over two rows	The rows are added, but only when that is what makes the figures add up to the total.	Tax-over-two-rows 77% to 100%
A unit column in the description; credit lines	A short unit ("pcs", "h") in its own column is not part of the description. Quantity is told from price by size, ignoring the minus sign of a credit.	Wide table 56% to 100%, credit lines 83% to 100%
A title with the reference beside it	A reference on the same row as a word like "Invoice" or "Faktura" is read as the invoice number.	Invoice number 39 to 50 of 50

3. Results on all 50 invoices

Field	Report 7	Report 8
Tax ID	29/50	50/50
Customer name	29/50	50/50
Tax amount	29/50	50/50
Subtotal	31/50	50/50
Tax rate	31/50	50/50
Total	31/50	50/50
Currency	36/50	50/50
Issue date	37/50	50/50
Due date	37/50	50/50
Invoice number	39/50	50/50
Shipping	42/50	48/50
Discount	46/50	50/50
Line description	782/813	813/813
Line quantity	788/813	813/813
Line unit price	788/813	813/813
Line amount	792/813	813/813
Supplier name	50/50	50/50

Page structure	Report 7	Report 8
dot leaders, no columns	11.2%	100.0%
wide table, item codes	56.1%	100.0%
pro forma with deposit	70.0%	100.0%
totals before items	72.7%	100.0%
tax over two rows	76.5%	100.0%
dated rows	78.1%	100.0%
credit lines	83.4%	100.0%
mixed tax rates	84.0%	98.9%
ruled grid	86.7%	100.0%
tear-off payment slip	86.7%	98.8%
bare	100.0%	100.0%
classic	100.0%	100.0%
coded	100.0%	100.0%

Page structure	Report 7	Report 8
inline	100.0%	100.0%
sidebar	100.0%	100.0%
stacked	100.0%	100.0%
wide	100.0%	100.0%

4. The older sets, scored again

The same extractor was run on the four older sets to check that the changes did not make them worse. The figures for corpora A, B and C come from my own runs of the saved extractor; corpus D was run by the same command as before.

Set	Field accuracy (Report 6 / now)	Mistakes found (Report 6 / now)	False alarms (Report 6 / now)
Corpus A: 30 independent invoices	99.01% / 99.01%	31 of 32 / 31 of 32	4 / 4
Corpus B: 40 generator invoices	99.83% / 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
Corpus D: 30 invoices	98.94% / 99.35%	26 of 26 / 26 of 26	8 / 4

Corpora A, B and C did not change, and corpus D improved a little as a side effect: its false alarms fell from 8 to 4, mostly because dates written out in Portuguese are now read.

5. What is still wrong

Two fields on the 50 invoices are wrong, both a shipping amount that is not on the invoice. Both invoices carry a planted subtotal mistake, so the figures cannot add up, and the extractor guesses that a leftover amount is shipping to make them do so. On corpus D, 8 fields of 1,098 are still wrong: a bare "\$" that could be one of several dollars, a day and month that cannot be told apart when every date has both numbers 12 or below, and a few labels that are printed in an unusual way.

The vision model was not run again. Its results in Report 7 stand.

Conclusion

The vocabulary, symbol and structure problems found in Report 7 are fixed in general ways. Field accuracy on corpus E went from 92.2% to 99.9%, and the ten invoices kept aside went from 64.9% to 100.0%, with no change to the older sets and a small gain on corpus D.

Read the numbers with care. The 50 invoices are synthetic, I wrote 30 of them and the extractor, and 40 of the 50 were used to make the fixes. The ten kept aside are few, and they share their ten layouts and ten countries with the others. A score on invoices in layouts and languages that were not part of this work would be the real test of the fixes. Chinese text is not scored.

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