KONICA MINOLTA

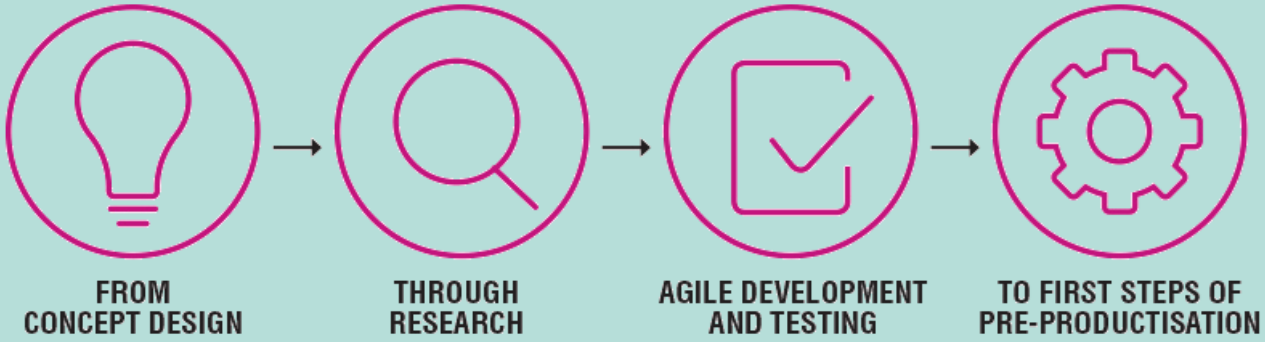
GLOBAL R&D
Research for new values

INVOICE MINING

NLPC Seminar

2022-04-20

Zuzana



Giving Shape to Ideas

INVOICE MINING

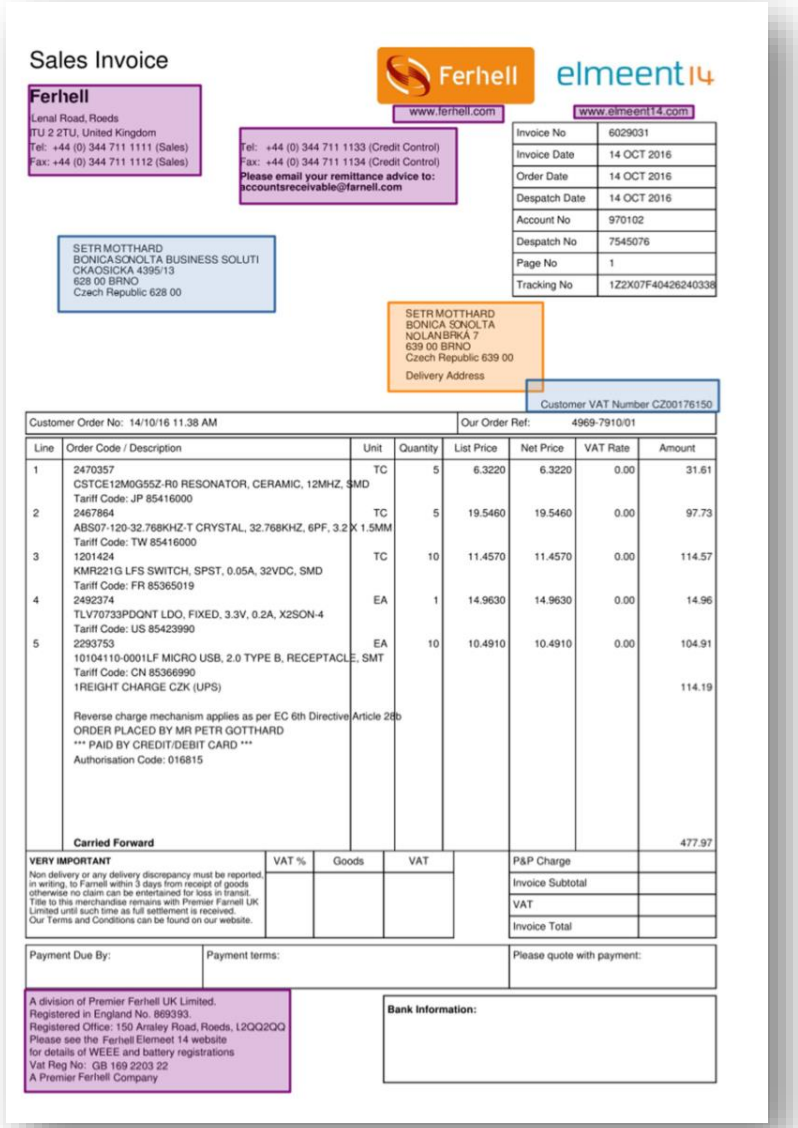
- User wants to enter data from incoming invoice into Accountant Information System (AIS) or ERP, either scanned or electronic
- User without access to AIS wants to overview invoices from a supplier or for a customer, and/or within a particular range of dates or amount

Pain points

- Ingesting the invoice into AIS, the user has to read and retype large parts of the invoice
- Without AIS, users have to open invoices one by one and extract the information manually

Pain reliever

- From electronic invoice or scanned invoice, extract information about buyer, seller, delivery address, amount, invoice date, due date, invoice number, and related order number
- Provide the extracted information either to the user or to other software (e.g. connector to AIS)



Sales Invoice

Ferhell
Lenal Road, Roeds
ITU 2 2TU, United Kingdom
Tel: +44 (0) 344 711 1111 (Sales)
Fax: +44 (0) 344 711 1112 (Sales)

www.ferhell.com

Tel: +44 (0) 344 711 1133 (Credit Control)
Fax: +44 (0) 344 711 1134 (Credit Control)
Please email your remittance advice to:
accountsreceivable@farnell.com

www.elmeent14.com

Invoice No	6029031
Invoice Date	14 OCT 2016
Order Date	14 OCT 2016
Despatch Date	14 OCT 2016
Account No	970102
Despatch No	7545076
Page No	1
Tracking No	1Z2X07F40426240338

SETR MOTTHARD
BONICA SONOLTA BUSINESS SOLUTI
CKAOSICKA 4395/13
628 00 BRNO
Czech Republic 628 00

SETR MOTTHARD
BONICA SONOLTA
NOLANBRKA 7
639 00 BRNO
Czech Republic 639 00
Delivery Address

Customer VAT Number CZ00176150

Customer Order No: 14/10/16 11.38 AM
Our Order Ref: 4969-7910/01

Line	Order Code / Description	Unit	Quantity	List Price	Net Price	VAT Rate	Amount
1	2470357 CSTCE12M0G55Z-R0 RESONATOR, CERAMIC, 12MHZ, SMD Tariff Code: JP 85416000	TC	5	6.3220	6.3220	0.00	31.61
2	2467864 ABS07-120-32.768KHZ-T CRYSTAL, 32.768KHZ, 6PF, 3.2 X 1.5MM Tariff Code: TW 85416000	TC	5	19.5460	19.5460	0.00	97.73
3	1201424 KMR221G LFS SWITCH, SPST, 0.05A, 32VDC, SMD Tariff Code: FR 85365019	TC	10	11.4570	11.4570	0.00	114.57
4	2492374 TLV70733PDQNT LDO, FIXED, 3.3V, 0.2A, X2SON-4 Tariff Code: US 85423990	EA	1	14.9630	14.9630	0.00	14.96
5	2293753 10104110-0001LF MICRO USB, 2.0 TYPE B, RECEPTACLE, SMT Tariff Code: CN 85366990	EA	10	10.4910	10.4910	0.00	104.91
	1REIGHT CHARGE CZK (UPS)						114.19
	Reverse charge mechanism applies as per EC 6th Directive Article 28b ORDER PLACED BY MR PETR GOTTHARD *** PAID BY CREDIT/DEBIT CARD *** Authorisation Code: 016815						
	Carried Forward						477.97

VERY IMPORTANT
Non delivery or any delivery discrepancy must be reported, in writing, to Farnell within 3 days from receipt of goods otherwise no claim can be entertained for loss in transit. Title to this merchandise remains with Premier Farnell UK Limited until such time as full settlement is received. Our Terms and Conditions can be found on our website.

VAT %	Goods	VAT

P&P Charge	
Invoice Subtotal	
VAT	
Invoice Total	

Payment Due By:	Payment terms:	Please quote with payment:

A division of Premier Farnell UK Limited.
Registered in England No. 869393.
Registered Office: 150 Arraley Road, Roeds, L2QQ2QQ
Please see the Farnell Elemeet 14 website
for details of WEEE and battery registrations
Vat Reg No: GB 169 2203 22
A Premier Farnell Company

Bank Information:

COMPONENTS

KONICA MINOLTA

Component	Description
OCR	Obtain text + coordinates from scanned document
OCR error correction	Based on external knowledge, resolve OCR imperfections
Keyword identification	Find keywords relevant for invoice and extraction language
Invoice block recognition	Split invoice into logical blocks: invoice info, buyer info, seller info, goods info etc. based on keywords + geometry
Block classification	Distinguish buyer from seller and from delivery address
Named entity recognition	Recognize words or expressions that represent person name, organization name, location or product name
Address recognition	Recognize parts of address: street, number, city, ZIP, etc.

WHAT INFORMATION TO EXTRACT FROM INVOICES?

KONICA MINOLTA

Balance feasibility & …

- Some information is quite easy to extract (e.g. dates)
- Some information is difficult to extract (e.g. complete company name)

… & usability

- Extraction of some fields has higher value than extraction of others

EXTRACT LANGUAGE-INDEPENDENT INFORMATION

Recognize the following fields

- Page number
- Invoice number
- Customer number
- Invoice date, due date
- Phone number, fax, email, web address
- Order number
- Bank account number
- VAT number
- Organization ID

Recognize and distinguish money amounts

- Unit price
- Product/service price
- Total amount (incl. VAT or excl. VAT)

- More difficult for OCR (vocabulary or language model does not help)
- Suitable for regular expressions

EXTRACT LANGUAGE-DEPENDENT INFORMATION

KONICA MINOLTA

Recognize the following fields:

- Invoice title
- Organization name
- Person name
- Address (street name + number, city, ZIP, country)
- Bank name
- Payment terms

- Depends on quality of NER for particular language.
- Suitable for (multiligual) NER, we use pretrained BERT
- Suitable for fine-tuning on own data

Distinguish the fields according to **context** into:

- Buyer information
- Seller information
- Delivery address information

HANDLE OCR/EXTRACTION ERRORS

Set format checks

- Normalize phone numbers, addresses, VAT numbers, dates, IBAN

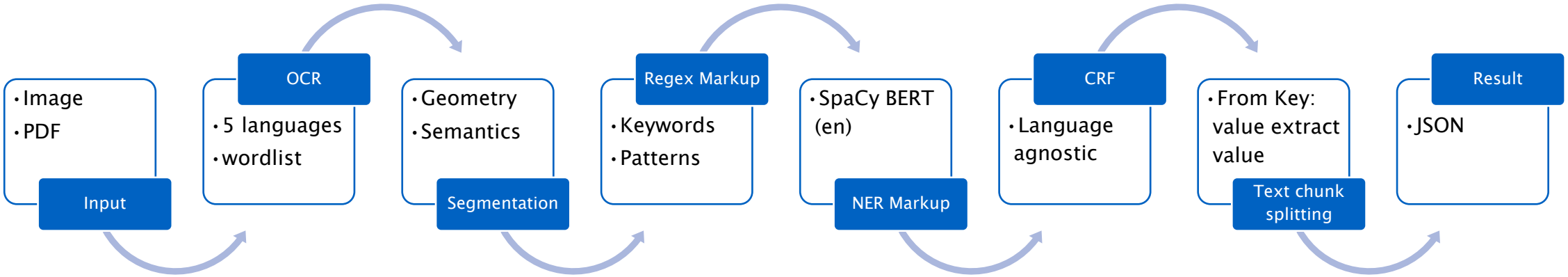
Perform fuzzy matching

- Solve OCR issues (common OCR errors: capital I = lowercase L)

Use customized models

- Wordlists
- Character lists
- Fine-tune OCR model on own data

PROCESSING PIPELINE



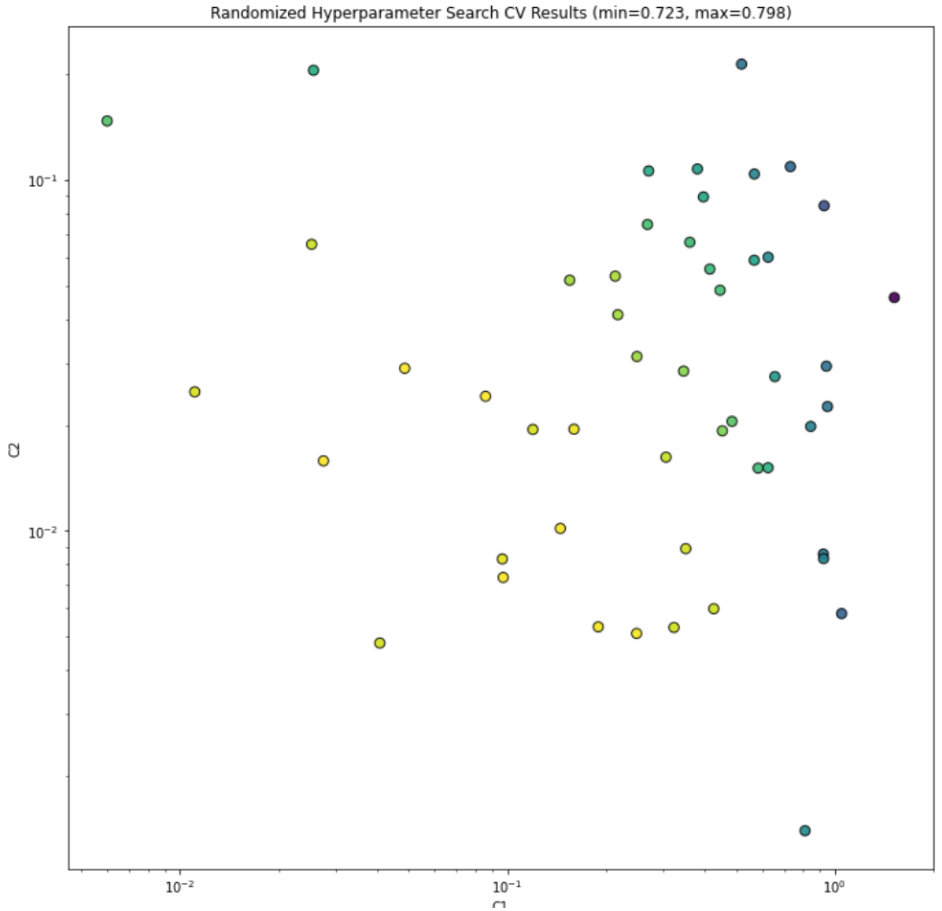
CRF CLASSIFIER

top	height	left	width	orig_text	right	unidecode	annotations
799	77	1867	287	INVOICE |	2154	invoice	invoice_title.key invoice_number.key
937	29	200	63	TO:	263	to	buyer.key
919	35	1083	48	ye	1131	ye	
987	29	202	208	Kristina Guiné	410	kristina guine	PERSON
1037	37	201	422	Germiphène Corporation	623	germiphene corporation	ORG
1087	29	1747	199	Invoice No .:	1946	invoice no	invoice_number.key
1087	29	2007	106	E0441	2113	e0441	general_identifier.value
1078	48	2171	12	|	2183	|	
1130	20	1342	28	'D	1370	d	
1105	65	1072	35	\	1107	\	
1228	48	1774	310	Date: July 6,2010 |	2084	date: july 6,2010	issue_date.keyvalue

data to features

```
{'left': False,
 'right': True,
 'top': True,
 'bottom': False,
 'farleft': False,
 'farright': False,
 'fartop': False,
 'farbottom': False,
 'block_num': 28,
 'small_length': True,
 'big_length': False,
 'text.isupper()': True,
 'text.istitle()': False,
 'text.isdigit()': False,
 'contains_space': True,
 'same_unidecode': False,
 'small_font': True,
 'big_font': False,
 'very_small_font': False,
 'very_big_font': False,
 'isKey': True,
 'isValue': False,
 'due_date.key': True,
 'payment_term.key': True,
 'text': 'DUE DATE’,
 'aligned_left_belowleft': False,
 'aligned_left_belowright': True,
…
}
```

CRF: OPTIMIZATION, EXPLANATION




```
Top positive:
10.052933 B-DATE   aligned_right_belowpage.value
7.897679 B-TO-PERSON aligned_left_abovevat_id.keyvalue
7.814449 I-TO-ORG aligned_right_belowvery_big_font
7.627294 I-TO-ORG aligned_left_abovedelivery.key
7.612133 I-TO-ORG aligned_line_leftinvoice_number.key
7.150480 B-FROM-ORG email.keyvalue
7.050193 I-TO-ORG web.keyvalue
7.012798 O        aligned_right_below4digit.value
6.895260 B-NUMBER aligned_left_belowissue_date.keyvalue
6.760817 O        aligned_left_below2account_number_nocode.value
6.723070 B-FROM-ORG aligned_right_abovevat_id.value
6.702354 B-FROM-PERSON aligned_right_below2email.value
6.693501 B-NUMBER aligned_left_abovepage.keyvalue
6.558869 I-FROM-ORG aligned_right_abovevat_id.key
6.532396 B-TO-ORG aligned_left_above2bottom
6.284452 O        aligned_left_below2iban.value
6.277688 B-FROM-ORG aligned_line_rightvat_id.keyvalue
6.202203 I-TO-ORG aligned_left_aboveprice.key
6.201795 I-NUMBER aligned_left_above2swift.keyvalue
6.175395 B-NUMBER aligned_line_leftvery_big_font
5.838829 B-NUMBER aligned_right_aboveaccount_number_with_code.value
5.778927 B-FROM-PERSON aligned_left_below2email.keyvalue
5.712573 B-TO-ORG aligned_right_beloworg_name.value
5.661203 B-TO-PERSON aligned_left_abovedelivery.key
```

INTERNAL EVALUATION – MUC5 METRICS

Evaluation scheme:

- Strict
- Loose

ACTual = COR + INC + PAR + SPU
POSsible = COR + INC + PAR + MIS
Precision = COR / ACT
Recall = COR / POS

Precision = (COR + PAR*0.5)/ ACT
Recall = (COR + PAR*0.5)/ POS

F1 = 2*Precision*Recall /
(Precision + Recall)

- **COR** – correct
- **INC** – incorrect
- **PAR** – partial
- **MIS** – missing
- **SPU** – spurious
- **NON** – noncommittal

	true	pred	eval
invoice_number	PDDD00-0000001	0036629276	INC
invoice_date	11/28/2016	None	MIS
seller_org	SAFETY-KLEEN SYSTEMS, INC	SAFETY-KLEEN SYSTEMS, INC	COR
seller_person	None	None	NON
buyer_org	Germiphene Corporation	Germiphene	PAR
buyer_person	None	Marc Detoa	SPU

EXTERNAL EVALUATION – PROPOSED METRIC

KONICA MINOLTA

EQUALITY	MUST HAVE fields	NICE TO HAVE fields	SUCCESS	FAILURE
Strict	Invoice number Issue date Amount	Due date VAT number Phone number Bank account number	Exact value (excl. spaces) extracted	Missing or incomplete value
Loose	Buyer org name Seller org name	Buyer person name Bank name	Spaces, person titles, “Ltd”, “Inc”, punctuation do not make a difference	>80% of words incorrect or missing

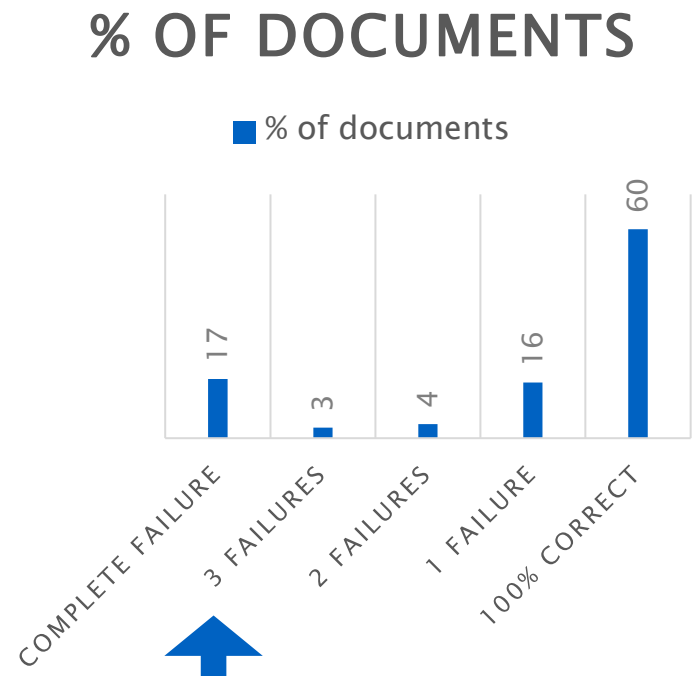
EXTERNAL EVALUATION – PROPOSED METRIC

EQUALITY	MUST HAVE fields
Strict	Invoice number Issue date Amount
Loose	Buyer org name Seller org name

In case of failure of MUST HAVE fields

Document extraction fails completely

In case of failure NICE TO HAVE fields consider the number of failures over which the document extraction fails completely.



The more fields we extract, the worse numbers.