A High Quality Data Pipeline for Reasonable-Scale Machine Learning

TAV 47, Google Munich, November 3rd 2022

David Faragó, Innoopract, EclipseSource Group

Overview

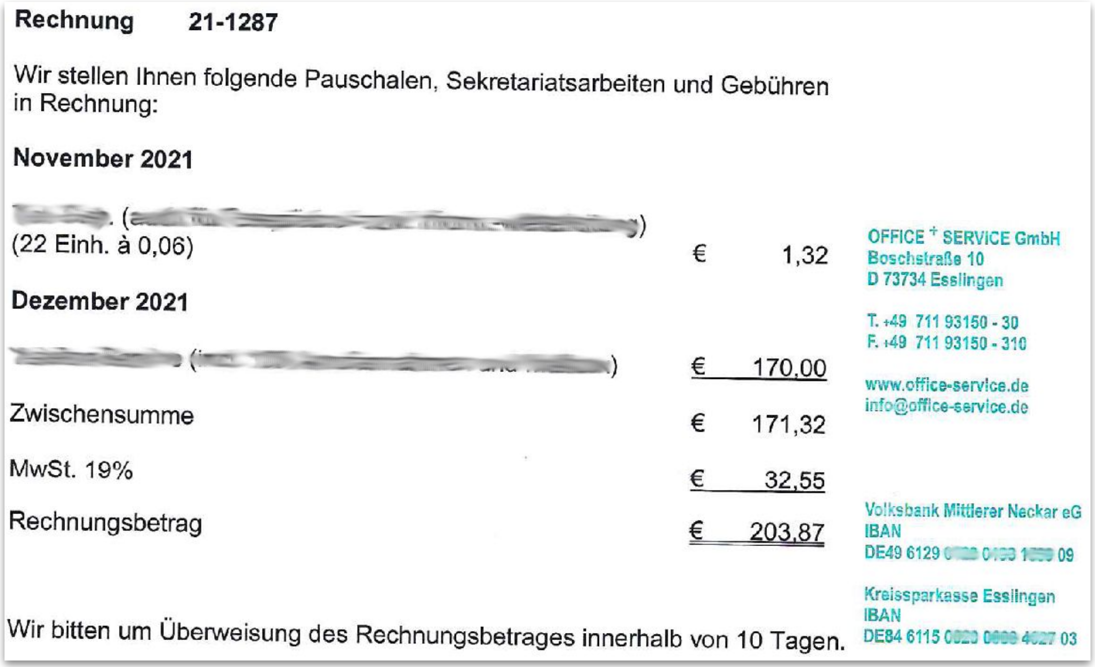
1) Data Quality

2) Pipeline to Generate Data

3) Pipeline Extension to Measure Data Quality

Setting

- reasonable scale
  - as little cost (people and hardware) as possible
  - focus on practice
  - open-source solutions
- 1 year of experience
- case study: KIE from invoices
  - most prominent IDP application
  - biggest challenge:
    high quality data[2]



Rechnung **21-1287**

Wir stellen Ihnen folgende Pauschalen, Sekretariatsarbeiten und Gebühren in Rechnung:

November 2021

[illegible]. ([illegible]) (22 Einh. à 0,06)	€	1,32
Dezember 2021		
[illegible] ([illegible])	€	170,00
Zwischensumme	€	171,32
MwSt. 19%	€	32,55
Rechnungsbetrag	€	203,87

Wir bitten um Überweisung des Rechnungsbetrages innerhalb von 10 Tagen.

OFFICE + SERVICE GmbH
Boschstraße 10
D 73734 Esslingen

T. +49 711 93150 - 30
F. +49 711 93150 - 310

www.office-service.de
info@office-service.de

Volksbank Mittlerer Neckar eG
IBAN
DE49 6129 [illegible] 09

Kreissparkasse Esslingen
IBAN
DE84 6115 [illegible] 03

Data Quality: Motivation

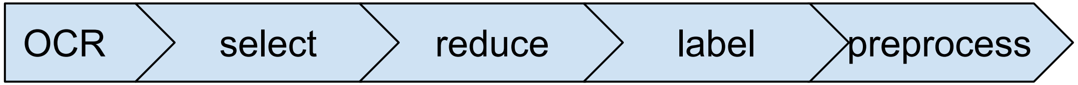
- often[18] “high quality” = high resolution images

- data-centric AI[7]
    - systematically engineer the data for AI
    - spend vast majority of time on data!
    - major movement in AI

Data Quality: Dimensions (Merged[1,3,5,17] & Extended)

Dimension	Definition	Quality degrading example
feature noise	percentage of incorrect feature values (wrt ground truth)	ground truth “50€” OCRed as “5i€”
class noise	feature noise on the target feature (i.e. the class)	a value incorrectly labeled as CUSTOMER-NO instead of INVOICE-NO
distribution noise	distribution distance between the dataset and the ground truth	training-serving skew with the training dataset being invoices from B2B, but serving for clients with B2C invoices
incompleteness	percentage of values missing	ground truth “50€” missing in the dataset due to OCR skipping it
inconsistency	percentage of values with more than one representation	TOTAL “50” and “50€” occurring in the dataset
redundancy	percentage of (non-exact) duplicates	two identical invoices, or with (almost) the same key information
class imbalance	average pairwise size difference between classes	most text of an invoice is no key information, leading to a much larger class OTHER

Core Data Pipeline



© 2022 EclipseSource | https://eclipsesource.com | David Faragó | A High Quality Data Pipeline for Reasonable-Scale Machine Learning

Core Data Pipeline Tasks and Technologies

Pipeline step	Task	Technologies
OCR	native mobile OCR and rotation	ML KIT (Android) & Vision (iPhone)
select	separate German from rest	Python Polyglot
	separate invoice types (giro, QR code)	CLIP, Pyton, Huggingface
reduce	variance-preserving size reduction	CLIP, Pyton, Huggingface
	remove invalid invoices	CLIP, Pyton, Huggingface
label	collaborative annotation guide	Google Docs
	label total/customer&invoice no/IBAN/...	Kotlin multiplatform, Jetpack Compose
preprocess	normalize	Python, Huggingface
	sanitize	Python, Huggingface
	abstract features (numbers, recipient)	Python, Huggingface
	tokenize	Python, Huggingface
	crop (for non-focus mode)	Python, Huggingface
	word embedding	Python, Keras or Fasttext

Exemplary Technology: Data Selection via CLIP

- OpenAI's Contrastive Language Image Pre-training (CLIP)[15]
  - similarity between images and captions
  - use similarity threshold to include / exclude images
  - iterative & semi-automatic since similarity thresholds vary strongly
  - semantically select ≈10% from 10^5 images into multiple, use-case specific datasets
- tasks
  - remove invalid invoices via similarity to the caption "Image of an invoice page containing a company name, an invoice number, a customer number, a total amount, and an IBAN."
  - separate invoice types (invoices with QR code / giro transfer) via similarity to a given image of that type (invoice with QR code / giro transfer)
  - reduce (non-exact) duplicates via pairwise image similarity

Exemplary Technology: Labeling Images

own labeling tool, specialized on Key-Information-Extraction from images

Data Pipeline, Including Quality Measurements

core data pipeline:

OCR | select | reduce | label | preprocess *

direct data quality check:

* measure

data quality check via ML model:

* split | augment | train | validate | test

Data Pipeline Quality Measurement Tasks and Technologies

Pipeline step	Task	Technologies
measure	statistics or review on dataset	e.g. Huggingface DataMeasurmentTool, Labeling Tool
	schema inference and validation	e.g. Great Expectations, Tensorflow Data Validation
split	train/valid/test split without data leakage	Pyton, Huggingface
augment	translate bounding boxes	Pyton, Huggingface
	shuffle words tagged OTHER	Pyton, Huggingface
	permute sequence order	Python, Huggingface
	oversample non-OTHER fields	Pyton, Huggingface
	vary number encoding	Python, Huggingface
train	train BiLSTM or Transformer model	Python, Tensorflow, Keras, Huggingface
validate	F½ model performance (boxes & fields)	Python, sklearn, seqeval, W&B
test	deploy on mobile device	Kotlin multiplatform, TFLite
	deploy in own labeling tool	Kotlin multiplatform, TFLite
	error analysis	Kotlin multiplatform, Jetpack Compose

Data Quality Measurements

Method	Quality Dimensions	Measurements from	Exemplary technology
manually	all but distribution noise	dataset	labeling or reviewing tool
comparison to given gold standard	all, mainly class noise	dataset	class noise interrater agreement[6,22]
statistics on datasets (distribution distance, outlier detection, …)	signals for all but inconsistency	dataset	Huggingface's Data Measurements Tool[8]
statistics via schemas (inferred or specified manually)	signals for all	datasets & schemas	Great Expectations[10], Tensorflow Data Validation[3]
model performance	signals for all, mainly class noise	the trained model the dataset was created for	see previous slides on data quality check via ML model
model confidence	signals for correctness dimensions	the trained model the dataset was created for	confidence learning tool Cleanlab[13]
predictions from quality prediction models	signals for correctness dimensions	quality prediction models	Consensus Filter[4, 19, 1]

Exemplary Task: Quality Measurement by Model Performance

- bad model performance ⇐ bad data quality
- model performance metric should be suitable for business case and risk: F½
- metric should be measured without data leakage
  - if your business case requires generalization to unseen invoice layouts
  - different recipient datasets for test, validation & train set (many papers[2] don't)
  - augment only on train set
- average F½ for our BiLSTM model:

field	F½	precision	recall
company name	0.82	0.84	0.76
invoice number	0.76	0.83	0.61
customer number	0.66	0.67	0.61
total	0.78	0.94	0.5
IBAN	0.97	0.98	0.93

- possible data issues: class imbalance, too little data (esp. customer number)

Exemplary Task: Manual Quality Review (60 invoices)

- difficult: redundancy (1 invoice), inconsistency, distribution noise, class imbalance

	COMPANY	TAG-INVOICE-NO	INVOICE-NO	CUSTOMER-NO	TAG-TOTAL	TOTAL	TAG-IBAN	IBAN	O
feature noise	10			1			16	11	?
class noise	12	1	1				2	1	?
incompleteness	3	1			2	1	2	2	?
inconsistency	49					1			?

- average: 2 issues/invoice. only 3 invoices with 0 issues
- low data quality for COMPANY due to logos, explains bad focus mode on logos
- 6 issues due to labeling mistakes, all other issues due to OCR
- otherwise minor and similar issues (not reflected by quality metrics)
- IBAN feature noise mitigated by post-processing, become inconsistency issues

Summary

- elaborate data pipeline, exemplified for KIE on images
- data quality measurements give insights (but find better quality metrics)
- DATA-CENTRIC AI !

© 2022 EclipseSource | https://eclipsesource.com | David Faragó | A High Quality Data Pipeline for Reasonable-Scale Machine Learning

References

[1] Al-Sabbagh, Khaled Walid et al. *Improving test case selection by handling class and attribute noise.* JSS 183. 2022

[2] Baviskar, Dipali et al. *Multi-Layout Unstructured Invoice Documents Dataset: A Dataset for Template-Free Invoice Processing and Its Evaluation Using AI Approaches*. IEEE Access, vol. 9. 2021.

[3] Breck, Eric, et al. *Data Validation for Machine Learning.* MLSys. 2019.

[4] Brodley, Carla, and Friedl, Mark. *Identifying and eliminating mislabeled training instances.* National Conference on AI. 1996.

[5] Budach, Lukas, et al. *The Effects of Data Quality on Machine Learning Performance.* arXiv preprint. 2022.

[6] Cohen, Jacob. *A coefficient of agreement for nominal scales.* EPM. 1960.

[7] *Data-centric AI Resource Hub:* https://datacentricai.org.

[8] https://huggingface.co/blog/data-measurements-tool.

[15] Radford, Alec, et al. *Learning transferable visual models from natural language supervision.* ICML. PMLR. 2021.

[17] Scannapieco, Monica. *Data Quality: Concepts, Methodologies and Techniques.* Data-Centric Systems and Applications. 2006.

[18] Scheuerman, Morgan Klaus et al. *Do datasets have politics? Disciplinary values in computer vision dataset development.* ACM on Human-Computer Interaction 5. 2021.

[19] Sluban, Borut, Gamberger, Dragan, and Lavra, Nada. *Advances in class noise detection.* ECAI. 2010.

[22] Barrett, Leslie, and Michael W. Sherman. *Improving ML Training Data with Gold-Standard Quality Metrics*. 2019.