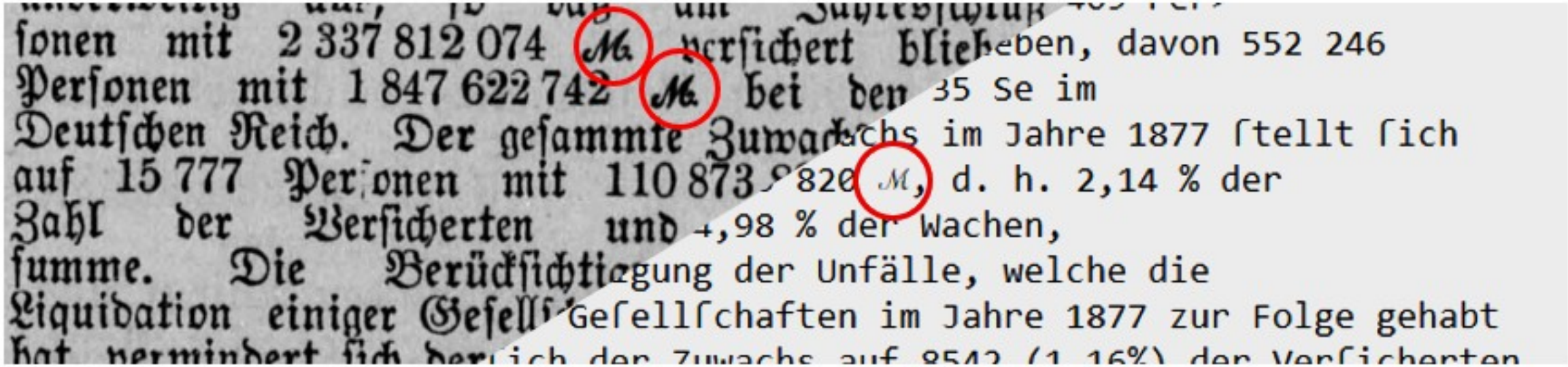
Ground Truth erstellen, OCR-Modelle verbessern



Stefan Weil

2022-10-20

OCR-D
Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)

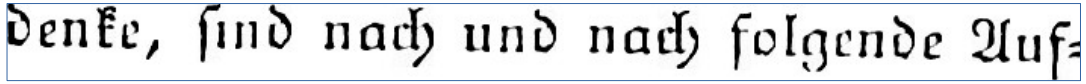
Gute OCR benötigt gute Modelle

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- OCR-Software nach heutigem Stand der Technik arbeitet mit neuronalen Netzen.
- Neuronale Netze müssen trainiert werden, damit sie bestmögliche Ergebnisse liefern.
- Trainingsergebnisse („Modelle“) werden für OCR-Software von OCR-D (Tesseract, Calamari, kraken) für die Texterkennung benötigt.

Training benötigt Ground Truth

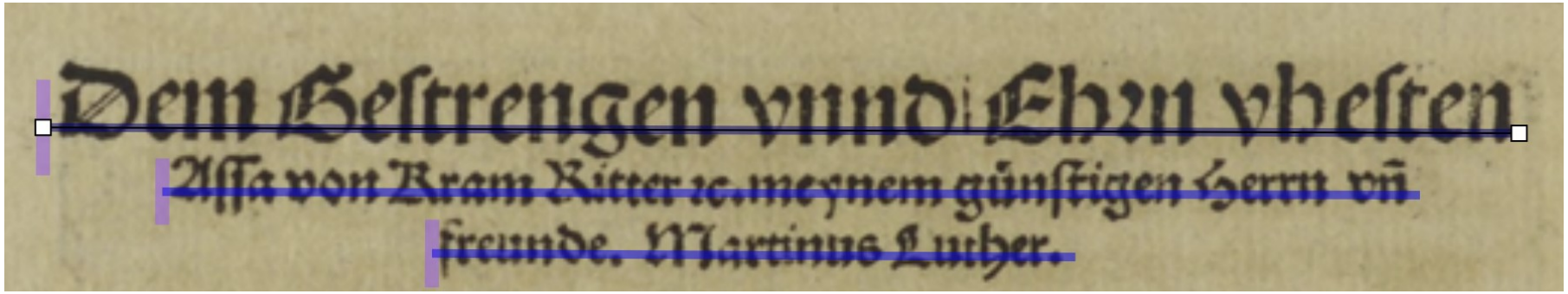
- Tesseract: Zeilenbild + Zeilentext („GT-Pärchen“)



denke, ſind nach und nach folgende Auf—

Problem / Fehler in GT Text: Trennstrich?

- Calamari, kraken: Seitenbild + Text mit Grundlinien (PAGE XML)



OCR-D Ground Truth Level

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- OCR-D definiert drei Level für die Transkription von Text: https://ocr-d.de/de/gt-guidelines/trans/
- UB Mannheim hat sich für Level 2 entschieden: langes s: ſ, rundes r: ꝛ, historische Umlaute: **aͤoͤ** usw., …
- Besonderheit: Identische Grapheme für I und J in Frakturschrift werden gemäß Lautwert transkribiert
- Vorhandene Ground Truth entspricht häufig nur Level 1

Ground Truth für Modelltraining an der UB Mannheim

Aufwertung vorhandener Ground Truth: Korrekturen, Umstellung von OCR-D GT Level 1 auf Level 2

- GT4HistOCR (Springmann, Reul, Dipper, Baiter)
  https://code.bib.uni-mannheim.de/ocr-d/GT4HistOCR
- AustrianNewspapers (Mühlberger, Hackl)
  https://github.com/UB-Mannheim/AustrianNewspapers/
- Neue Zürcher Zeitung (Ströbel, Clematide)
  https://github.com/UB-Mannheim/NZZ-black-letter-ground-truth

Ground Truth für Modelltraining an der UB Mannheim

Ground Truth der UB Mannheim

- Fibeln, Weisthuemer, digi-gt, digitue-gt,
  alle auf GitHub bei
  https://github.com/UB-Mannheim/
- reichsanzeiger-gt
  https://github.com/UB-Mannheim/reichsanzeiger-gt/
  (wird auch für Training der Layouterkennung verwendet)

Umfang Ground Truth für historische Zeitungen

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Reichsanzeiger: 96k Zeilen, 420k Wörter (aktueller Stand, perspektivisch ca. 110k Zeilen und 440k Wörter),
  viele kurze Zeilen durch transkribierte Tabellen
- Austrian Newspapers: 58k Zeilen, 325k Wörter
- Neue Zürcher Zeitung: 43k Zeilen, 300k Wörter

Modelltraining

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- 2019: Tesseract-Modell auf Basis von GT4HistOCR
  https://github.com/tesseract-ocr/tesstrain/wiki/GT4HistOCR
- 2022: Modelle für Tesseract, Calamari und kraken auf Basis von AustrianNewspapers
  https://github.com/UB-Mannheim/AustrianNewspapers/wiki
- 2022: weitere Modelle für Tesseract und kraken
  https://ub-backup.bib.uni-mannheim.de/~stweil/tesstrain/

Nachtraining

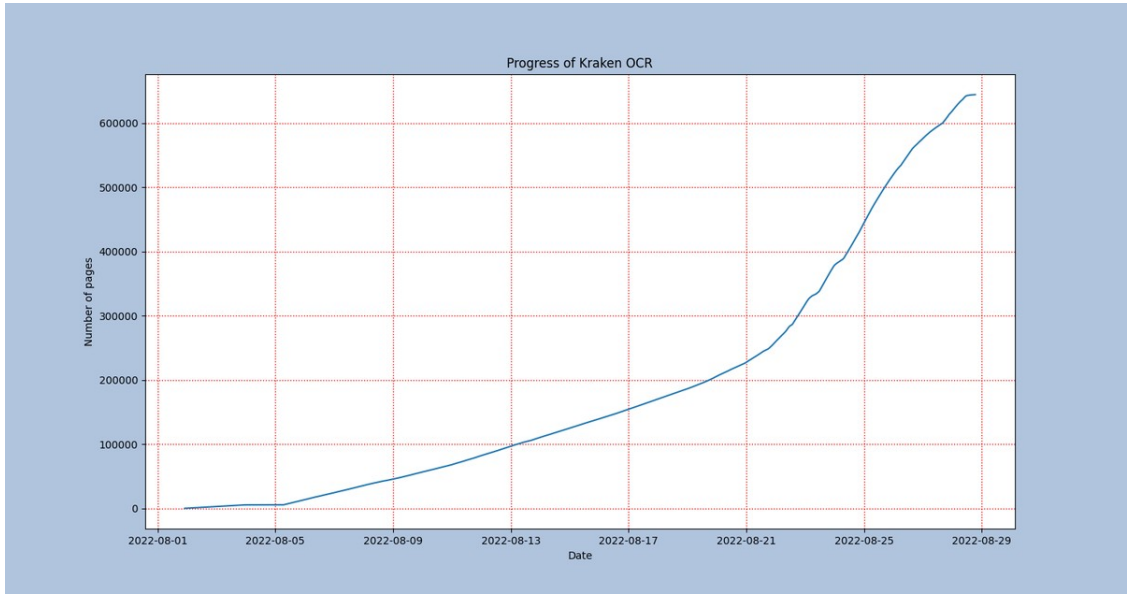
- Vorhandene Modelle lassen sich um neue Glyphen erweitern oder für bestimmte Schriften optimieren.
- Beispiel: Modelle für kraken unter Verwendung der Ground Truth von AustrianNewspapers, digi-gt, reichsanzeiger-gt, digitue-gt und Trainingsmaterial für Schreibmaschinenschrift

Modelle

austriannewspapers
↓
luther (aus digi)
↓
reichsanzeiger
↓
digitue
↓
typewriter

Praktische Anwendung

- Erzeugung von Volltexten mit neuen kraken-Modellen für rund 1400 Werke mit mehr als 650000 Seiten im Rahmen des Projektes OCR-BW (UB Mannheim, UB Tübingen, 2019–2022)

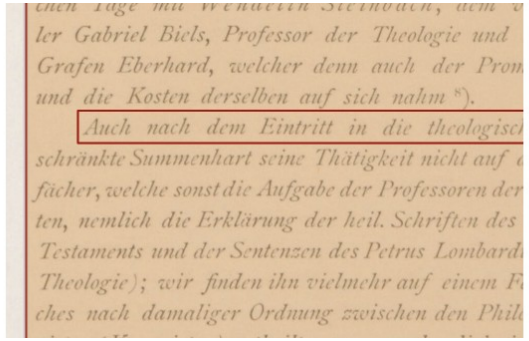


https://github.com/UB-Mannheim/digitue-gt/wiki

Demonstration Nachtraining

Verbesserung der Volltexte durch Nachtraining

vorher



anwalnte

und die Kosten derselhen at sich naline [2]).

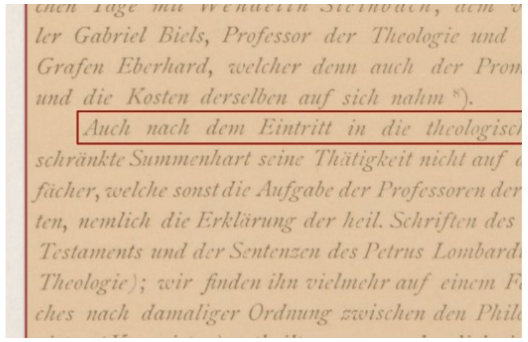
Auch uach dem Linfritt in die Meolggische lacnttat be-

ränhte Jummcnhart seine Thatigtcit nicht ai diefnigen

Lchr≠

acher, welcle voustdie Lcgabe der Bolessoren der

nachher



anwohnte

und die Kosten derselben auf sich nahm [2]).

Auch nach dem Eintritt in die theologische Facultät be-

schränkte Summenhart seine Thätigkeit nicht auf

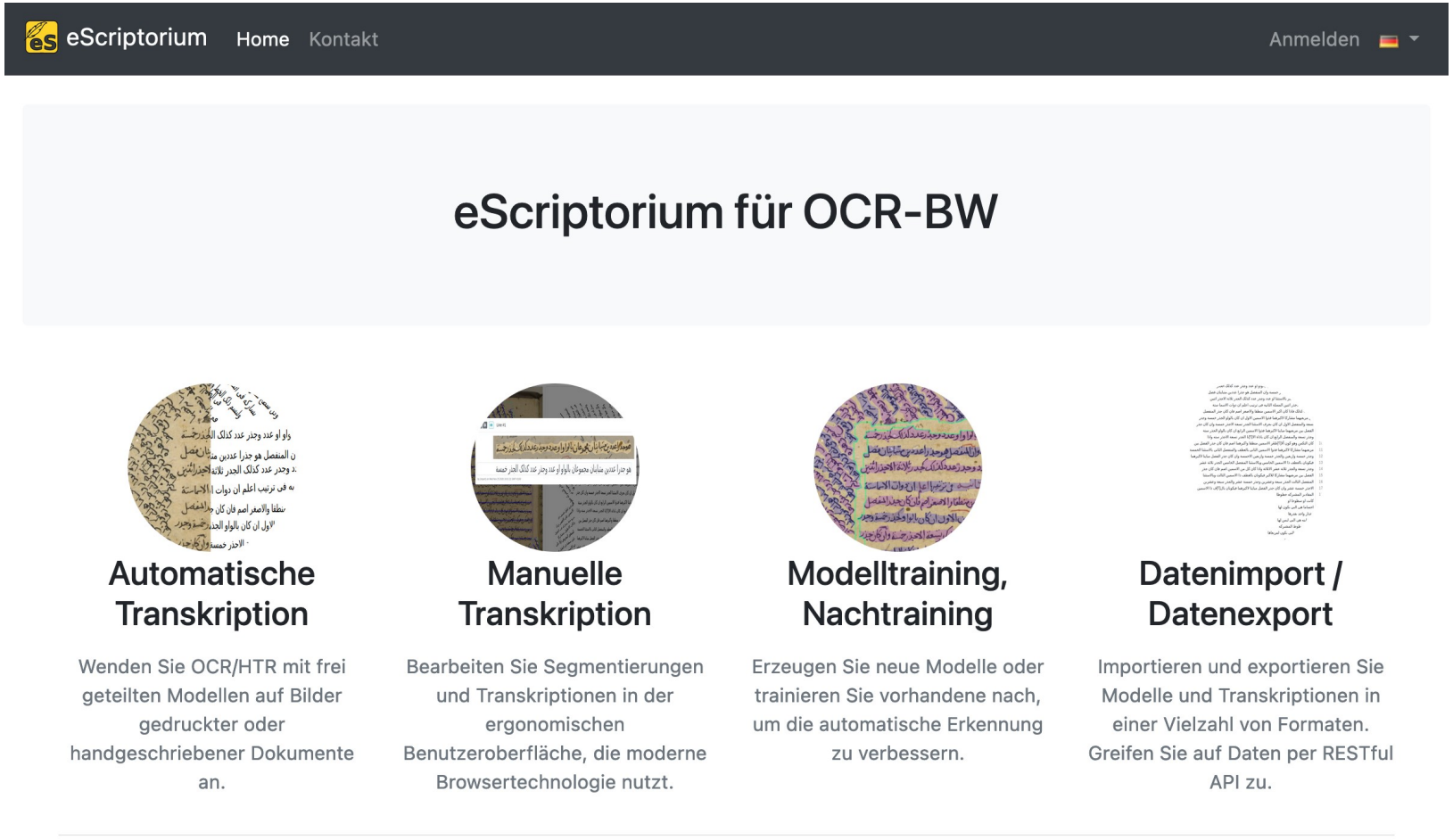
diejenigen Lehr-

fächer, welche sonst die Aufgabe der Professoren der

Erkannte Probleme

- Neue „Umlautvarianten“:
  Ground Truth enthält unterschiedliche Transkriptionen für historische Umlaute → neuronales Netz kann sich nicht für eine Variante entscheiden, sondern nimmt beide. Vermutlich wäre es sinnvoll, Umlaute und andere diakritische Zeichen als Kombination eines normalen Zeichens mit einem kombinierenden Zeichen (Unicode: combining character) zu trainieren.
- Außerdem etliche typische Buchstabenverwechslungen, die sich auch in der Ground Truth als Fehler wiederfinden
- Garbage in → garbage out.
  Ground Truth muss für bessere Modelle weiter verbessert werden.

eScriptorium als Werkzeug für GT-Erstellung und Nachtraining



Workflow mit eScriptorium

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Import eines Werkes (Seitenbilder) in Web-Applikation eScriptorium, z. B. per IIIF-Manifest
- Automatisierte Layouterkennung und Transkription mit Hilfe ausgewählter kraken-Modelle
- Manuelle Korrektur von Layouterkennung und Transkription für ausgewählte Seiten
- Nachtraining der Modelle für Layout- und Texterkennung in eScriptorium
- Export der neuen Modelle für produktiven Einsatz
- Export der GT als PAGE XML für Publikation (z. B. auf GitHub)

Ausblick

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Ein erster Prototyp von eScriptorium kann jetzt auch Modelle von Tesseract und Calamari für die Texterkennung verwenden.
- Und auch das Modelltraining funktioniert nicht mehr nur für kraken, sondern auch für Tesseract und Calamari.

Damit wird die maschinelle Texterkennung einschließlich Nachtraining auch ohne OCR-Expertenwissen möglich.

Literatur

- Springmann, Uwe, et al. Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. arXiv preprint arXiv:1809.05501 (2018).
- Ströbel, Phillip and Clematide, Simon. Improving OCR of Black Letter in Historical Newspapers: The Unreasonable Effectiveness of HTR Models on Low-Resolution Images (2019). In: Proceedings of the Digital Humanities 2019 (DH2019)
- Weil, Stefan. Neue Frakturmodelle für Tesseract. Kitodo-Anwendertreffen (2019). https://madoc.bib.uni-mannheim.de/53748
- Mühlberger, Günter and Hackl, Günter. NewsEye / READ OCR training dataset from Austrian Newspapers (19th C.) (2019). Zenodo. http://doi.org/10.5281/zenodo.3387369
- Kamlah, Jan and Schmidt, Thomas. Finetune your OCR! Improving automated text recognition for early printed works by finetuning existing Tesseract models. ELAG (2022).