OCR-D:
Was wurde bisher erreicht und wie geht es jetzt weiter?

Clemens Neudecker (@cneudecker)
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

DHd AG Zeitungen & Zeitschriften

OCR - Herausforderungen und Lösungen für Zeitungen & Zeitschriften

Frankfurt am Main, 11. November 2019

Einleitung

- Massendigitalisierung in Bibliotheken hat inzwischen Millionen von Dokumenten digital als Scans verfügbar gemacht, aber wenig Volltexte
- Trotz vieler Projekte in der Vergangenheit (z.B. IMPACT, eMOP, etc.) noch immer unzureichende OCR Qualität historischer Drucke für die Nachnutzung bspw. in den Digital Humanities
- Erhebliche Fortschritte im Bereich Document Analysis und Recognition durch Verwendung von Deep Learning Verfahren (RNN, CNN, LSTM)
- Deep Learning Verfahren bzw. Modelle erfordern spezifische Anpassungen auf die Bedarfe von Bibliotheken und die Besonderheiten historischer Dokumente

→ DFG Ausschreibung am 28. Mai 2014

OCR-D Koordinierungsprojekt

- Seit 2015: DFG-Förderung für OCR-D Koordinierungsprojekt
- Koordinierungsprojekt mit 4 Partnern:
  - Herzog August Bibliothek Wolfenbüttel
  - Berlin-Brandenburgische Akademie der Wissenschaften
  - Staatsbibliothek zu Berlin (ab 12/2016)
  - Karlsruher Institut für Technologie (ab 08/2017)
- Kernaufgaben des Koordinierungsprojekts:
  - Handlungsbedarfe identifizieren und systematisieren (Phase I)
  - Technische Anforderungen und Rahmenbedingungen für OCR-Entwicklung spezifizieren (Phase I)
  - Betreuung von OCR-D Softwareentwicklungs-Modulprojekten und Teststellung der OCR-D Software in Pilotbibliotheken (Phase II)

OCR-D Spezifikationen (1/2)

- Konsequent offene und Community-basierte Entwicklung via GitHub
  - `OCR-D` auf GitHub: https://github.com/OCR-D | https://ocr-d.de/
  - `Chat`: https://gitter.im/OCR-D/Lobby

- METS Container (basierend auf den Anforderungen der DFG-Praxisrichtlinien und des DFG-Viewer)
  - `METS` in OCR-D: https://ocr-d.de/mets

- PAGE-XML für OCR Ergebnisse (perspektivisch mit Transformationsszenarien nach ALTO, TEI)
  - `PAGE-XML` in OCR-D: https://ocr-d.de/page

- Taverna Workflow Engine für Prozessketten und Provenance
  - `Taverna` in OCR-D: https://github.com/OCR-D/taverna_workflow

OCR-D Spezifikationen (2/2)

- Kommandozeile als minimale Anforderung für OCR-D Software:
  - `CLI`: https://ocr-d.de/cli
- JSON Schema für OCR-D Softwarebeschreibung:
  - `ocrd-tool.json`: https://ocr-d.de/ocrd_tool
- ZIP+BagIt für (Ground Truth) Daten:
  - `OCRD-ZIP`: https://ocr-d.de/ocrd_zip
- Docker als Container für OCR-D Software:
  - `Docker`: https://ocr-d.de/docker

OCR-D Referenzimplementierung

- Konsequente Verwendung von `Python3` für alle OCR-D Software (wenn immer möglich)
- `core` Referenzimplementierung unterstützt Software-Entwickler und Anwender mit
  - `ocrd_utils` = logging, path normalization, coordinate calculation etc.
  - `ocrd_models` = file format wrappers for PAGE-XML, METS, EXIF etc.
  - `ocrd_modelfactory` = instantiate models from existing data
  - `ocrd_validators` = validating BagIt, ocrd-tool.json, METS, PAGE, CLI
- https://github.com/OCR-D/core oder https://pypi.org/project/ocrd/
- API docs https://ocr-d.de/core/

OCR-D Ground Truth Repository & Daten

- Erstellung von detaillierten Transkriptionsrichtlinien für Ground Truth Daten mit PAGE-XML
  - https://ocr-d.de/gt
- Entwicklung eines Repository für Ground Truth Daten
  - https://github.com/OCR-D/repository_metastore
- Bereitstellung von Ground Truth Daten aus OCR-D
  - https://ocr-d.de/gt-repo
- Semantisches Labeling von Ground Truth Daten
  - https://github.com/OCR-D/gt-labelling

OCR-D Modulprojekte

- Seit 2017 OCR-D Phase II mit 8 eigenständigen und durch die DFG geförderten OCR-D Modulprojekten:
  - [MP1] Bildoptimierung (DFKI Kaiserslautern)
  - [MP2] Layouterkennung (DFKI Kaiserslautern)
  - [MP3] Layouterkennung (Uni Würzburg)
  - [MP4] Nachkorrektur (Uni Leipzig)
  - [MP5] Optimierung von Tesseract OCR (UB Mannheim)
  - [MP6] Nachkorrektur (Uni München)
  - [MP7] Schriftarterkennung und Trainingsinfrastruktur (Uni Erlangen, Mainz, Leipzig)
  - [MP8] Langzeitarchivierung (SUB Göttingen)
- https://ocr-d.de/projects

Weitere OCR-D Komponenten

- Zusätzlich Bereitstellung von externer Software mit OCR-D Schnittstellen-konformität durch OCR-D Koordinierungsprojekt um Lücken in den Modulprojekten zu schließen bzw. für Vergleiche mit SoTA:
  - `ocrd_calamari`
  - `ocrd_im6convert`
  - `ocrd_kraken`
  - `ocrd_ocropy`
  - `ocrd_olena`
  - `ocrd_segment`
  - `ocrd_kerasLM`
  - `dinglehopper`
- http://kba.cloud/ocrd-kwalitee/
- `ocrd_train`: Makefile zum Trainieren von Tesseract LSTM

→ „Adoptiert“ von Tesseract als https://github.com/tesseract-ocr/tesstrain

OCR-D Einstieg und eigenes Experimentieren

- OCR-D Tutorial der DHd2019
  - http://kba.cloud/2019-03-25-dhd/
- OCR-D Setup Guide
  - https://ocr-d.de/docs/setup-2019-10-27
- OCR-D Chat
  - https://gitter.im/OCR-D/Lobby

Sonderfall Zeitungen (& Zeitschriften)

- Qua Beauftragung liegt das Hauptaugenmerk von OCR-D auf den VD-Digitalisierungsprojekten (VD16, VD17, VD18)

→ Keine Arbeiten/Anpassungen in OCR-D spezifisch für Zeitungen & Zeitschriften!

- Besondere Herausforderungen bei Zeitungen:
  - Mehrspaltiges Layout
  - Kleine bzw. stark variierende Schriftgrößen
  - Komplexe Reihenfolge von Regionen (Artikelseparierung)
  - Hoher Anteil an nicht-textuellen Regionen (Bilder, Tabellen, Werbung)
  - Niedrige Qualität der Vorlagen bzw. Digitalisate (Papier, Mikrofilm)

Ausblick

- Aktuell größtes Desiderat: Spezielle Layouterkennung für (historische) Zeitungen & Zeitschriften
  - Sind die Textregionen erst ordentlich segmentiert, Verwendung der regulären OCR-D Softwarekomponenten
  - Aktuelle Arbeiten an der SBB im QURATOR Projekt (Code coming soon!)

- Idee: Kombination von Layouterkennung (optische Merkmale) mit NLP Methoden – z.B. Transformer – (sprachliche Merkmale) um Reihenfolge der Regionen zu bestimmen

- Ground Truth bzw. Trainingsdaten werden benötigt!

Danke für die Aufmerksamkeit!
Fragen?

Clemens Neudecker (@cneudecker)
Staatsbibliothek zu Berlin – Preußischer Kulturbesitz

DHd AG Zeitungen & Zeitschriften

OCR - Herausforderungen und Lösungen für Zeitungen & Zeitschriften

Frankfurt am Main, 11. November 2019