Vom Bild zum Text – praktische OCR für die DH

—

15.09.2021, 14–16 Uhr: Abschlussveranstaltung

Eventreihe

- Dienstag, 23.03.2021, 10–12 Uhr: Einführungsveranstaltung
- Mittwoch, 05.05.2021, 15–17 Uhr: OCR-D, OCR4all, TEI-Konvertierung
- Mittwoch, 12.05.2021, 15–17 Uhr: Evaluation, Transkription, Training
- Mittwoch, 19.05.2021, 15–17 Uhr: Postcorrection, Hackathon
- **Mittwoch, 15.09.2021, 14–16 Uhr: Abschlussveranstaltung**

Agenda

- Recap und jüngste Entwicklungen
- Ideen für die weitere Planungen
- Wo geht es weiter?

Recap und jüngste Entwicklungen

- Preprocessing: Deskewing, Despeckling, Dewarping, Binarisierung
- Segmentierung: Erkennung von Regionen, Zeilen, Marginalien, Überschriften, Lesereihenfolge, ...
- Texterkennung: die "eigentliche" OCR
- Evaluation und Nachkorrektur: Erkennen und ggf. korrigieren von OCR-Fehlern (aber nicht historische Schreibweisen!)

OCR-D & OCR4all

- Ziel: Volltextdigitalisierung der VD-Bestände (Massenverarbeitung)
- Kommandozeilentools
- Breite Abdeckung von state-of-the-art Werkzeugen
- Vollautomatische Workflows

https://ocr-d.de

- Ziel: Erkennung einzelner Werke (Fehlerfreiheit)
- Grafische Oberfläche
- Fokus auf ausgewählte Werkzeuge mit robuster Standardkonfiguration
- Semi-automatische bzw. manuelle Workflows

https://ocr4all.org

Evaluation, Transkription, Training

- Evaluation: Wie gut ist die Text- und Layouterkennung
- Transkription: Erstellen von Trainingsdaten für spezifische OCR-Modelle
- Training: Erstellen von Modellen, die auf ein Werk oder eine Domäne zugeschnittene OCR-Modelle erlauben

Jüngste Entwicklungen

- Kick-off der dritten Projektphase von OCR-D am 30. Juli 2021
  - 4 Implementierungsprojekte
  - 3 Modulprojekte
  - ocr4all-libraries (für bestmögliche Nutzung von OCR-D Lösungen in OCR4all)
- LAREX Version 0.6.0
  - (fast) volle OCR-D Kompatibilität
  - “echter” XML-Editor (Tags, die in LAREX nicht editiert werden können, bleiben beim Speichern erhalten)
  - Diff-View zwischen GT und Prediction
  - viele Kleinigkeiten
- Calamari Version 2.1.3
  - umfassend konfigurierbare Regularisierungen
  - schneller und besser :-)

Ideen und weitere Planungen

Verbesserung der Layoutanalyse

- Noch nicht zufriedenstellend gelöst
- Positive Entwicklungen, speziell in Form trainierbarer (!) Methoden
- Idee: Kombination von Crowdsourcing und Active Learning
  - Nutzer:innen wenden existierende Lösungen auf eigenes Material an
  - Identifikation und **gezielte** Korrektur von Problemfällen
  - Einspeisen der Korrekturen in das Trainingskorpus
  - Neutraining der Modelle
  - Und von vorne
- Datenverwaltung, Training, etc. weitestgehend zentralisiert
- Unterstützung durch DHd Verband möglich/wahrscheinlich

Transkribathon?

- Vorstellen von und üben mit Transkriptionsumgebungen
- Klärung von Fragen (Kodierung, Transkriptionsrichtlinien, …)
- Nutzer:innen bearbeiten betreut Daten
  - eigene?
  - bislang unterrepräsentierte?
  - ggf. analog zur Layoutanalyse?
  - …?
- Training von werk- oder domänenspezifischen Modellen
- Teilen von GT und Modellen?
- …?

AG Kolloquium

- 2020 kein AG Treffen in persona
- Ersatz: “Weihnachts- und Neujahrskolloquium - Einblicke und Ausblicke”
  - AG-Mitglieder geben Einblicke in ihre Arbeiten (OCR-Engines, Editionen, Handschriftenerkennung, ...)
  - Offen und flexibel
  - Mehrere kurze (1-2h) Termine statt einzelner, ganztägiger Veranstaltung
- Erste Ausgabe 2020/21 wurde sehr gut angenommen
- Zweite Ausgabe in Planung
  - Gerne Input “von außen”! → mehr Infos demnächst über die Mailingliste
  - Gerne Alltagsprobleme anstatt “Cutting Edge Research“

Feedbackrunde

- Konnten Sie die Inhalte der vDHd-Reihe für Ihre eigenen OCR-Projekte nutzen?
- Hätten Sie sich noch weitere Inhalte in der Reihe gewünscht?
- Wünschen Sie ggf. noch weitergehenden Austausch?
- Haben Sie Interesse an den vorgestellten Projekten/Veranstaltungen teilzunehmen?
- Haben Sie noch weitere Vorschläge für AG-Projekte o.ä.?
- ...?

GT- und Modellrepo

– bestehende Lösungen

– Probleme

– Nutzerwünsche

– wie organisieren?

Wo geht es weiter?

Kommunikationskanäle

- DHd AG OCR Website
- DHd AG OCR Mailingliste
- Zweiwöchentlich: OCR-D Open Tech Call (Fokus auf Entwicklung und Technologien)
- Zweiwöchentlich: OCR-D GT Call (Fokus auf Ground Truth, Transkription und Training)
- Monatlich: OCR-D & Co (Barcamp zu OCR-Themen, Teilnehmer entscheiden Themen, niedrigschwellig)
- Chat with us:
  - https://gitter.im/dhd-ag-ocr/community
  - https://gitter.im/OCR-D/Lobby

OCR-Community gesucht

OCRd, 4all, AG-, Gitter, Veranstaltungen, Kooperationen, …