Qualitätssicherung für massenhafte Volltexterzeugung mit OCR-D

Absichtserklärung zur Antragstellung innerhalb der DFG-Ausschreibung „Implementierung der OCR-D-Software zur Volltextdigitalisierung“ vom 27. Februar 2020

Antragsteller

1. Universität Leipzig, Abteilung für Sprachverarbeitung
   vertreten durch Prof. Dr. Gerhard Heyer
2. Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden
   vertreten durch den Generaldirektor Dr. Achim Bonte

Projektziele

An beiden beteiligten Institutionen wurden im Rahmen von OCR-D-Projektphase 2 verschiedene **Vorarbeiten** geleistet. Im Modulprojekt „Unsupervised OCR-Postcorrection based on Neural Networks and Finite-state Transducers“ wurden an der ASV Leipzig u.a. Prozessoren für Nachkorrektur, OCR-Evaluierung und Bewertung von Text mittels Sprachmodellen entwickelt, sowie Bildvorverarbeitung, Segmentierung und Texterkennung (als funktional unverzichtbare Teile des Gesamtsystems). Außerdem wurden mit den vorhandenen Prozessoren in mannigfaltigen Workflows auf dem bereitgestellten Ground-Truth umfassende Messungen zur erreichbaren Qualität durchgeführt. Während der Pilotphase an der SLUB Dresden wurden zahlreiche Prozessoren getestet und mitentwickelt. Für einen größeren Dokumentenbestand, das *Börsenblatt für den Deutschen Buchhandel*, wurde ein geeigneter Workflow erarbeitet und durchprozessiert.

Diese intensiven Tests haben gezeigt, dass die **Varianz** in der Qualität des erzeugten Volltextes in Abhängigkeit des gewählten Workflows und des zu prozessierenden Dokuments sehr hoch ist: Die geeignete Auswahl der OCR-D-Prozessoren und ihrer Parametrisierung ist ein entscheidendes Kriterium für eine erfolgreiche Volltexterstellung. Aktuell fehlt es im OCR-D-Kontext sowohl an *Best Practices*, die helfen, für ein gegebenes Dokument und dessen

Metadaten einen passenden Workflow zu definieren, als auch an Methoden und Metriken, um zu überprüfen, ob der Workflow zu einem plausiblen Ergebnis geführt hat. Beide Aspekte sind vor dem Hintergrund der massenhaften Volltexterzeugung, wie sie in wissenschaftlichen Bibliotheken durchgeführt werden, von entscheidender Bedeutung für die Akzeptanz der OCR-D-Werkzeuge.

An dieser Stelle setzt das vorgeschlagene Projekt an: Wir planen die Entwicklung von Heuristiken und Modellen, die auf Ebene der im OCR-D-Funktionsmodell definierten abstrakten Prozessorklassen jeweils eine automatische **Abschätzung der Güte** des Prozessschrittes auf dem konkreten Digitalisat ermöglichen, und welche die **Konfiguration** von konkreten Prozessoren und ihrer Parameter zu einem geeigneten Workflow unterstützen. Diese Heuristiken und Modelle vergleichen zwischen dem jeweils erreichten Teilergebnis und einem anhand manuell erzeugter Trainingsdaten (sog. *Ground-Truth*) vorab ermittelten impliziten Erwartungswert. Hier findet also (anders als bei der Überprüfung der generellen Leistungsfähigkeit einzelner Methoden oder Module, wie sie bspw. in wissenschaftlichen Wettbewerben zum Einsatz kommt) bewusst kein Vergleich mit dokumentspezifischem Ground-Truth statt. Die Evaluierung basiert auf erlerntem, generalisiertem Vorwissen, welches sich ggf. dokumentklassenspezifisch (d.h. mit Hilfe von Metadaten) faktorisieren lässt.

Als **Datengrundlage** und **Bestandsziel** zugleich sehen wir das *Deutsche Textarchiv* (DTA) vor. Das DTA bietet mit seinen hochwertigen, manuell erfassten und nachkontrollierten Volltexten einen in Qualität und Quantität einzigartigen Datenbestand, der die für das Vorhaben nötige Stratifikation in Entstehungszeit und inhaltlicher Varianz sicherstellt und für die von OCR-D primär angestrebten Bestände der im deutschen Sprachraum erschienen Drucke des 16., 17. und 18. Jahrhunderts hinreichend repräsentativ ist. Bisher ist es nicht gelungen, diesen Datenschatz für die Nachnutzung im OCR-D-Kontext aufzubereiten.

Diese Entwicklung mündet in folgende **Deliverables**:

1. Auf Basis des DTA ensteht ein normierter, hochqualitativer, OCR-D-konformer Ground-Truth für Dokument- und Seitenstruktur sowie Textinhalt. Die fehlenden Strukturdaten werden über den Umweg von Struktur- und Texterkennung mit Hilfe der vorhandenen Textrepräsentation automatisch rekonstruiert.
2. Definierte Gütekriterien geben dem Nutzer bzw. dem Workflow-Managementsystem ein nachvollziehbares **Feedback** über Erfolg oder Nicht-Erfolg der Prozessierung bzw. einzelner Prozesschritte. Im Problemfall kann mindestens frühzeitig abgebrochen werden. Sofern entsprechende alternative Prozessierungsoptionen konfiguriert sind, besteht außerdem die Möglichkeit zur automatischen Wiederholung des jeweiligen Prozessschritts mit anderen Parametern oder in einer anderen Implementierung. So können ggf. ganze Prozessketten im Workflow ersetzt werden (dynamische **Reprozessierung** zur Laufzeit).
3. Generelle Empfehlungen (**Best Practices** zu statischer Workflow-Konfiguration) helfen dem Anwender bei der Identifikation geeigneter

Workflows: Welche Art von Dokument benötigt welche Prozessoren mit welchen Parametern für welchen Prozessschritt? Diese Empfehlungen werden abgeleitet aus den bei der Ermittlung von Erwartungswerten am Ground-Truth zu sammelnden systematischen Messungen zu der erreichbaren Qualität (in unterschiedlichen Workflow-Konfigurationen auf den verschiedenen Dokument- und Seiten-Klassen). Neben der einmaligen gesamtheitlichen Auswertung gehört dazu integral die **Visualisierung** einzelner Einflussfaktoren von Prozess- und Datenseite.

4. In Abhängigkeit von der vorhandenen Datengrundlage wird mit einer **Ablationsstudie** untersucht, wie stark mit zunehmender Größe zufällig gewählter Teilmengen der Lernstichprobe die jeweiligen Qualitätsmaße konvergieren, um auch eine Abschätzung über die Unsicherheit und **Generalisierbarkeit** der Beobachtungen treffen zu können (ohne bereits explorativ weiteren Ground-Truth produzieren zu müssen).

Die Anforderungen bei der Entwicklung von Heuristiken und Modellen sind für die drei groben **Teilbereiche** des OCR-D-Funktionsmodells – Bildvorverarbeitung, Layoutanalyse und Texterkennung – sehr verschieden: Für die Abschätzung der Qualität von **Text** kann vorab auf unabhängige Ressourcen an Daten zurückgegriffen werden – rein textueller Ground-Truth ist in großer Menge und Vielfalt verfügbar –, woraus sich Sprachmodelle und Wörterbücher ableiten lassen. Bei **Segmentierung** und (Segment-)Klassifizierung dagegen müssen neuartige Modelle zur Qualitätsabschätzung entwickelt werden, die den vorhandenen Ground-Truth intelligent und ökonomisch zu nutzen verstehen. In der **Bildvorverarbeitung** wiederum ist kaum scharf zwischen der Abschätzung der Qualität der Bildvorverarbeitung und der eigentlichen Bildqualität zu unterscheiden. Auch gibt es in diesem Teilbereich weniger Möglichkeiten der Abstraktion und damit Modellbildung, sodass man mehr auf Heuristiken angewiesen ist. Generell gilt, dass die Qualität der vorgelagerten Prozessschritte für das in den jeweils nachfolgenden Arbeitsschritten Erreichbare kritisch ist. Es kommt darauf an, die Güte der einzelnen Prozessschritte mit ihrem differentiellen Beitrag zum Gesamtergebnis zu wichten.

Das Projekt begleitet also den gesamten Weg vom Digitalisat bis zum Volltext und ebnet den OCR-D-Werkzeugen den Weg von der Einzelanwendung in die produktive Massendigitalisierung.

Grobe Abschätzung des Mittelbedarfs

Die Projektpartner rechnen mit einem Gesamtaufwand von 48 Personenmonaten in der Personalkostenkategorie *Postdoktorandin/Postdoktorand und Vergleichbare* über 24 Monate Laufzeit zuzüglich Sachmitteln in Höhe von ca. 10 000 EUR für Reise- und Veranstaltungskosten.

22. Mai 2020

Prof. Dr. Gerhard Heyer

Dr. Achim Bonte