Zentrum für Philologie und Digitalität
„Kallimachos“
Geschäftsstelle
Lehrstuhl für Künstliche Intelligenz
und Wissenssysteme
Am Hubland, D-97074 Würzburg
Der Geschäftsführende Vorstand
UNIV.-PROF. DR. ULRICH KONRAD
Telefon 0931 / 31-82828
Telefax 0931 / 31-82830
ulrich.konrad@uni-wuerzburg.de

20. Mai 2020

Absichtserklärung zur Ausschreibung *Implementierung der OCR-D-Software zur Volltextdigitalisierung* im Rahmen des Förderprogramms *e-Research-Technologien*

Anpassung des Open-Source Tools OCR4all zur Gewährleistung der idealen Unterstützung existierender OCR-D Lösungen

Dieser Antrag im Rahmen der Ausschreibung zur dritten Phase des OCR-D Projekts hat zum Ziel, das GUI-basierte Open-Source Werkzeug OCR4all[1] so zu erweitern und anzupassen, dass die Anwendung existierender OCR-D Lösungen möglichst ideal unterstützt wird. Dadurch werden zwei Hauptziele verfolgt: Zum einen soll v. a. nicht-technischen Nutzern, die den Umgang mit der Kommandozeile nicht gewohnt sind, ein komfortabler Zugang zu OCR-D Lösungen zur Verfügung gestellt werden. Zum anderen unterstützt eine zusätzliche, visuelle Erklärungskomponente bei der Zusammensetzung und Konfiguration „optimaler“ Workflowlösungen für unterschiedliche Materialien.

Die geplante Antragstellung erfolgt durch das an der Julius-Maximilians-Universität Würzburg etablierte Zentrum für Philologie und Digitalität „Kallimachos“ (ZPD; Kommissarischer Geschäftsführender Vorstand Univ.-Prof. Dr. Ulrich Konrad)[2], vertreten durch den kommissarischen Leiter der Digitalisierungseinheit Christian Reul.

Projektziele und Anwendungsszenario

Das Hauptziel des Antrags ist es, OCR4all so anzupassen, dass die im Rahmen von OCR-D entwickelten Lösungen möglichst ideal angewendet werden können. Dabei soll der Fokus auf der einfachen und uneingeschränkten (Qualifikation der Nutzer, verwendetes Betriebssystem, Material, …) Anwendung auf üblichen Arbeitsplatzrechnern liegen.

[1] http://www.ocr4all.org
[2] https://www.uni-wuerzburg.de/zpd

Primär soll es damit einer deutlich größeren Nutzergruppe als bisher ermöglicht werden, OCR-D Lösungen selbstständig, sicher und komfortabel zur Massendigitalisierung einzusetzen, da sich gezeigt hat, dass die exklusive Nutzung der Kommandozeile für nicht-technische Nutzer durchaus eine nicht unerhebliche Hürde darstellt. Zusätzlich gehen wir allerdings fest davon aus, dass auch technisch versierten Nutzern, für die die Bedienung mittels der Kommandozeile kein Problem darstellt, speziell durch die erweiterten Möglichkeiten der Visualisierung und Interaktivität ein echter Mehrwert geboten werden wird:

Durch die enorme Heterogenität an zu verarbeitenden Materialien ist nicht zu erwarten, dass es die eine, globale Ideallösung geben wird, sowohl in Hinblick auf die Verarbeitungspipeline (Kombination von Prozessoren) sowie die verwendeten Settings (Parameter der Prozessoren und genutzte Modelle), die für jeden Anwendungsfall die optimalen Ergebnisse liefert. Es gilt also herausfinden, welche der zahlreichen und flexibel einsetzbaren OCR-D Lösungen für welche Aufgaben am besten geeignet sind. Um dies zu erreichen, ist eine rein quantitative Bewertung (erzielte Zeichenfehlerrate etc.), gerade auf anspruchsvollen historischen Drucken, nicht ausreichend. Für eine gezielte Optimierung muss zunächst ein Verständnis geschaffen werden, an welcher Stelle der Pipeline überhaupt eingegriffen werden muss und auf welche Weise der Eingriff erfolgen soll. Hierzu wird eine zusätzliche, visuelle Erklärungskomponente benötigt, die sowohl das Endergebnis als auch sämtliche relevanten Zwischenergebnisse intuitiv verständlich darstellen kann.

Um die beiden Hauptziele zu erreichen, sind, neben dem Schaffen der technischen Voraussetzungen (Adaption von Schnittstellen, Verfügbarmachung einzelner Prozessoren, …) einige tiefgehende konzeptionelle Anpassungen sowie umfangreiche Änderungen an der OCR4all Software nötig. U. a. soll ein Evaluationsframework bereitgestellt werden, in dem beliebige Pipelines aus verschiedenen Prozessoren und unterschiedlichen Parametern zunächst, auch von nicht-technischen Nutzern, komfortabel definiert und nachfolgend automatisch auf ausgewählten Testdaten angewendet werden können. Anschließend erfolgt ein strukturierter Vergleich der Ergebnisse, sowohl durch die erwähnte visuelle Erklärungskomponente als auch quantitativ, falls entsprechende Ground Truth vorliegt.

Des Weiteren muss der Anwendungsbereich von Einzelwerkebene auf eine beliebige Menge an Werken umgestellt werden. Diese sollen ggf. zu nutzerdefinierten Werkclustern mit ähnlichen Charakteristiken zusammengefasst werden können und anschließend mit den zuvor identifizierten „optimalen“ Settings vollautomatisch prozessiert werden.

Bestehende Vorarbeiten und Pilotierungsphase

Um die zuvor definierten Ziele zu erreichen, bauen wir auf das an der Universität Würzburg entwickelte und dauerhaft fest im ZPD verankerte Open-Source Werkzeug OCR4all auf, das komfortabel mithilfe einer übersichtlichen grafischen Weboberfläche bedient und mittels Docker oder VirtualBox einfach und unabhängig vom Betriebssystem eingerichtet und ausgeführt werden kann. Ziel ist es, auch explizit nicht-technischen Nutzern eine Möglichkeit zu geben, selbst die ältesten und anspruchsvollsten gedruckten Werke eigenständig, mit überschaubarem Zeitaufwand und in höchster Qualität zu erfassen.

Um die einzelnen Workflowschritte abzudecken, fokussiert sich OCR4all derzeit auf einige wenige, ausgewählte externe Open-Source Tools und Eigenentwicklungen: Neben dem bekannten OCRopus (Vorverarbeitung und Zeilensegmentierung) und dem im Projekt entstandenen LAREX (Segmentierung und interaktive Nachkorrektur) wurde auch die ebenfalls an der Universität Würzburg entwickelte OCR Engine Calamari bereits vollständig integriert. Die Ergebnisse fast jedes Workflowschritts können von den Nutzern einfach

und doch präzise nachkorrigiert werden. Neben der Minimierung von Folgefehlern kann dadurch ein (nahezu) fehlerfreies Ergebnis erreicht werden, was für manche Anwendungsfälle, z. B. digitale Editionen, unerlässlich ist. Die durch diese Korrekturen generierte Ground Truth wird wiederum zum Training spezialisierter OCR Modelle verwendet, um die Erkennungsrate weiter zu optimieren.

OCR4all wurde sowohl innerhalb der Universität Würzburg im Rahmen unterschiedlicher Forschungsvorhaben als auch auf nationaler wie internationaler Ebene hervorragend aufgenommen. Es findet mittlerweile auf einem vielfältigen Spektrum an Materialien nahezu jeden Alters durch eine fachwissenschaftlich sehr heterogene Nutzerbasis Anwendung und wird für verschiedenste Sprachen und Sprachstufen eingesetzt, um durch manuelle Eingriffe in einen semi-automatischen Prozess hochqualitative Ergebnisse zu erzielen. Dies wurde bereits umfangreich evaluiert[3].

In der dem Antrag vorangehenden Pilotierungsphase ist geplant, den existierenden OCR4all Workflow unter Verwendung der entsprechenden OCR-D Prozessoren prototypisch nachzubilden und dadurch einen lauffähigen Proof-of-Concept zu erstellen. Anschließend sollen unter Verwendung dieses Prototyps verschiedene deutschsprachige Werke des 16. bis 19. Jahrhunderts erfasst werden. Neben dem Sammeln von Erfahrungen und Nutzerfeedback, sowie dem Erkennen möglicher Schwachstellen, soll dabei außerdem für ausgewählte Werke Ground Truth erstellt werden, die dann wiederum für die Evaluation der übrigen OCR-D Prozessoren hinsichtlich Stabilität, Laufzeit, Performanz, etc. außerhalb von OCR4all verwendet wird.

Univ.-Prof. Dr. Ulrich Konrad
Geschäftsführender Vorstand des ZPD

[3] Vgl. Reul et al.: OCR4all - An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings *in Applied Sciences (2019). 9(22).* https://www.mdpi.com/2076-3417/9/22/4853/htm