OCR4all – Eine semi-automatische Open-Source-Software für die OCR historischer Drucke

Wehner, Maximilian
maximilian.wehner@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Dahnke, Michael
michael.dahnke@uni-siegen.de
Universität Siegen, Deutschland

Landes, Florian
florian.landes@kbl.badw.de
Bayerische Akademie der Wissenschaften, Deutschland

Nasarek, Robert
robert.nasarek@geschichte.uni-halle.de
Martin-Luther-Universität Halle-Wittenberg, Deutschland

Reul, Christian
christian.reul@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Das Problemfeld der OCR früher Drucke

Lange galt die automatisierte Texterkennung oder sog. Optical Character Recognition (OCR) historischer Drucke des späten Mittelalters und der Frühen Neuzeit, das heißt die Überführung des gedruckten Textes in eine maschinenverarbeitbare Form, als sehr problematisch (Rydberg-Cox 2009). Die OCR moderner Texte wird dagegen auch aufgrund technischer Innovationen wie des zeilen- statt zeichenbasierten OCR-Ansatzes (Breuel et al. 2013) weithin als informatisch gelöstes Problem angesehen. Die teils höchst komplexen Layoutstrukturen von Inkunabeln und der bis zum Ende des 18. Jahrhunderts gedruckten Werke, ihr oft schlechter Erhaltungs- und Druckzustand sowie die Vielfalt und Varianz der in ihnen verwendeten Drucktypen stellen dagegen bis heute sogar den kommerziellen State of the Art der Texterkennungssoftware wie beispielsweise ABBYY FineReader[1] vor erhebliche Probleme. Auch die vermeintlich einfach gedruckten Frakturromane des 19. Jahrhunderts bereiten bei ihrer Überführung in eine E-Text-Variante immer wieder große Schwierigkeiten. Trotz der durch Bibliotheken und andere öffentliche Einrichtungen bereit gestellten, wachsenden Bestände bilddigitalisierter Vorlagen dieser Epochen ist darum der Umfang digitalisierter Texte nicht annähernd im selben Maß gewachsen, obwohl in den vergangenen Jahren bereits deutliche Fortschritte für die OCR vormoderner Drucke aufgezeigt werden konnten (Springmann / Lüdeling 2017).

Vor allem für die geistes- und kulturwissenschaftliche Editionsphilologie eröffnet sich auf diese Weise ein erhebliches Problemfeld, ist diese vor dem Hintergrund der Entwicklung hin zu immer mehr digitalen Editionen doch auf meist große Textmengen in digitaler Form angewiesen, die im besten Fall neben ihrer hohen Zeichengenauigkeit bereits Metainformationen über das gedruckte Ursprungsmedium aufweisen – zu denken wäre hier besonders an die Typisierung unterschiedlicher Layoutregionen (Überschriften, Marginalien, Bildbeischriften etc.) oder die Lesereihenfolge der einzelnen Layoutelemente des originalen Textes. Und auch mit Blick auf neuere Forschungsfelder innerhalb der Geisteswissenschaften und Digital Humanities (Text Mining, Sentiment Analysis usw.) sowie deren Bedarf an großen Textmengen zur Anwendung quantitativer Analyseverfahren stellt sich zunehmend die Frage nach Möglichkeiten einer OCR früher und vormoderner Drucke, die sowohl hohen Qualitätsansprüchen als auch einem entsprechenden Automatisierungsgrad genügt.

Werkzeuge, die diese Anforderungen erfüllen, sollten zudem frei verfügbar sein, sich einfach und selbstständig von einem informatisch nicht vorgeschulten Nutzerkreis auf einer einheitlichen Benutzeroberfläche bedienen lassen und die unterschiedlichen Submodule wie beispielsweise die Vorverarbeitung von Bilddateien, Möglichkeiten der Layouttypisierung sowie die eigentliche Zeichenerkennung integrativ zu einem kohärenten OCR-Workflow zusammenführen.

Am Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik der Julius-Maximilians-Universität Würzburg wurde deshalb die OCR-Software OCR4all[2] entwickelt, welche die genannten Notwendigkeiten in sich vereint und sich als erstes Programm überhaupt mit Blick auf die besonders herausfordernden Textgruppen direkt an Geisteswissenschaftler*innen richtet.

OCR-Workflow

Typischerweise gliedert sich ein OCR-Workflow in vier Hauptkomponenten (s. Abbildung 1). Im sog. **Preprocessing** werden die Originalbilder in Vorbereitung späterer Arbeitsschritte binarisiert (Konvertierung des Ausgangsbildes in ein Schwarzweißbild) und gerade gestellt, um die nachfolgenden Arbeitsschritte zu erleichtern.

Abbildung 1: Hauptkomponenten eines typischen OCR-Workflows. Von links nach rechts: Originalbild, Preprocessing, Segmentierung, OCR, Nachkorrektur.

Während der **Segmentierung** erfolgt die Erkennung und Typisierung der Layoutbestandteile. Dazu werden zuerst die Text- und Nicht-Textregionen (Bilder, Bordüren etc.) unterschieden, optional die Textregionen anschließend als Haupttext, Überschriften, Marginalien etc. semantisch ausgezeichnet. Abschließend werden die Textregionen zur Vorbereitung der OCR in einzelne Zeilenbilder zerschnitten.

In einem dritten Schritt, der **OCR**, werden die identifizierten Bildzeilen durch die Anwendung von sog. Modellen in maschinenverarbeitbaren Text umgewandelt. Je nach Material können dazu entweder sog. gemischte Modelle verwendet werden, die mithilfe einer Vielzahl ganz unterschiedlicher, jedoch epochentypischer Werke erstellt wurden. Handelt es sich bei den zu bearbeitenden Werken hinsichtlich der Vielfalt und Varianz der in ihnen verwendeten Drucktypen sowie deren Erhaltungszustand jedoch um sehr spezifische Drucke, können sog. werkspezifische Modelle für die Erkennung erstellt und verwendet werden.

In der **Nachkorrektur** können die generierten maschinenverarbeitbaren Texte und Daten abschließend nachbearbeitet und korrigiert werden.

OCR4all orientiert sich in seinem Aufbau an den beschriebenen Hauptkomponenten eines OCR-Workflows, gliedert diese jedoch noch einmal in unterschiedliche Teilmodule. Der modulare Aufbau erlaubt dabei eine Einbindung und Verwendung bereits bestehender Softwarelösungen, die gemäß ihrer Stärken zu einem kohärenten OCR-Workflow kombiniert werden.

Grundsätzlich kann der Workflow vollautomatisch durchlaufen werden. Dennoch hat der Nutzer immer die Möglichkeit, korrigierend in jeden Teilschritt einzugreifen, um ein optimales Ergebnis zu garantieren, welches als Startpunkt des dann folgenden Teilschritts fungiert. Dafür können die für jedes Teilmodul vorgegebenen Einstellungen durch den Nutzer individuell angepasst werden.

Das Preprocessing erfolgt in OCR4all wie oben beschrieben. Dabei werden alle gängigen Eingabeformate für Bilddateien unterstützt. Dem schließt sich die Layouttypisierung mithilfe des Segmentierungstools LAREX[3] (s. Abbildung 2) an. Hier können werkspezifische Parameter zur Text- und Bildtypisierung festgelegt sowie zu erkennende Layoutregionen (Haupttext, Überschriften, Marginalien, Seitenzahlen etc.) definiert werden. Je nach Komplexität des vorliegenden Seitenlayouts ist nach einer automatischen Layouterkennung ein Eingriff in das vorliegende Ergebnis mittels unterschiedlicher Korrekturwerkzeuge möglich. Weiterhin kann in LAREX die Lesereihenfolge der Layoutbestandteile markiert werden, um den Lesefluss des Originals vorlagengetreu nachbilden zu können. Vor allem für die Verwendung des maschinenverarbeitbaren Textes in digitalen Editionen sind diese Funktionen unverzichtbar.

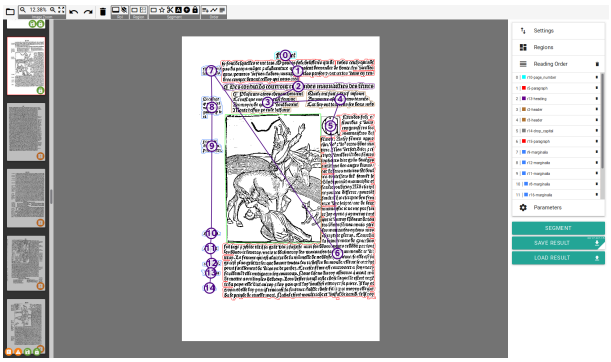

Abbildung 2: Im Teilmodul der Segmentierung erfolgen die Typisierung der Layoutelemente sowie die Festlegung der Lesereihenfolge.

Der Layouttypisierung folgt die Zeilensegmentierung. In dieser werden die Text beinhaltenden Layoutbestandteile in einzelne Zeilenbilder zerteilt (OCRopus[4]), um die eigentliche OCR vorzubereiten.

Anschließend wird im Erkennungsschritt aus den vorliegenden Einzelzeilen (mittels Calamari[5]) maschinenverarbeitbarer Text generiert. Dazu können in OCR4all bereits standardmäßig integrierte gemischte Modelle für Fraktur- und Antiquaschriften unterschiedlicher Epochen genutzt werden. Es besteht die Möglichkeit, die entstandenen Texte anschließend in einem Editor komfortabel zu korrigieren (s. Abbildung 3).

Abbildung 3: Im Editor kann generierter Text mithilfe eines sog. Virtual Keyboard (rechts) zeichengetreu korrigiert werden.

Für die Feststellung der Fehlerrate der Zeichenerkennung kann im Evaluationsmodul der ursprünglich erkannte Text mit der durch den Nutzer vorgenommenen Korrektur verglichen werden.

Darüber hinaus bietet OCR4all die Möglichkeit, die oben angesprochenen werkspezifischen Modelle unter Verwendung vorgenommener Textkorrekturen selbst zu trainieren, stetig zu verfeinern und anzuwenden. Besonders bei Werken mit erheblicher Typenvielfalt und -varianz, bei denen ein bestehendes gemischtes Modell keine hinreichenden Erkennungsergebnisse erzielt, können auf diese Weise dennoch sehr hohe Zeichenerkennungsraten erreicht werden.

In der abschließenden Nachkorrektur können die generierten Texte editionsreif korrigiert und als Plain Text oder PageXML[6] ausgegeben werden. Letzteres Format beinhaltet neben dem eigentlichen Text auch dessen Verankerung in semantischen Positionen auf den Druckseiten in Form von Koordinaten.

In Abhängigkeit des Ausgangsmaterials variiert der zum Erreichen einer sehr hohen Genauigkeit benötigte Arbeitsaufwand zwischen wenigen Minuten bei Werken mit einfachen Layoutstrukturen, für die ein passendes Modell vorliegt, und einigen Stunden bei sehr komplexen, frühen Drucken, für die werkspezifische Modelle trainiert werden müssen (Reul et al. 2019).

Workshopkonzeption

Der ganztägige Workshop soll einem informatisch und technisch nicht spezifisch vorgeschulten Nutzerkreis einen nachvollziehbaren und verständlichen Einstieg in das Themen- und Problemfeld der OCR historischer Drucke bieten. Er wird dazu befähigen, mithilfe der vorgestellten Software eigenständig qualitativ hochwertige Texte aus ganz unterschiedlich anspruchsvollen Ausgangsdaten zu generieren – und dies mit zeitlich vertretbarem Aufwand. Die Konzeption erfolgt aus diesem Grund sehr praxisbezogen. Konkret bedeutet dies einen angeleiteten und individuell betreuten Durchgang durch den oben vorgestellten OCR-Workflow anhand verschiedener, nach Layoutkomplexität, Typographie, Erhaltungszustand und Entstehungszeitraum geclusterter Drucke. Dabei sollen anwendungsbezogen wichtige Grundfragen der OCR beantwortet werden:

- Wie verändert sich entsprechend des Ausgangsmaterials die Anwendung der OCR-Workflows und der in ihm enthaltenen Submodule?
- Mit welchem Aufwand ist in unterschiedlichen Bearbeitungsphasen des Materials zu rechnen?
- Wie stark lässt sich der Workflow in Abhängigkeit des vorliegenden Materials automatisieren?
- Wie schnell sind bei einem werkspezifischen Training welche Erkennungsraten erreichbar?
- Welcher Aufwand ist mit Blick auf die spätere Verwendung der produzierten Texte überhaupt sinnvoll?
- …

Da sich neben den oben beschriebenen, meist vormodernen Textspezifika auch eine grundlegende technische Expertise der Benutzer*innen im Bereich der OCR als eine wichtige Bedingung für die Produktion hochwertiger digitaler Texte herausgestellt hat, strebt der Workshop neben einer besonders praktischen Handlungsanleitung auch die Vermittlung der wichtigsten Funktionskonzepte der in OCR4all integrierten Submodule an.

Der Workshop umfasst neben den oben beschriebenen Inhalten auch Fragen der Einrichtung und Installation der Software. Zusätzlich wird eine Serverversion der Software zur Verfügung gestellt, die einen reibungslosen Ablauf gewährleistet und Trainingsprozesse werkspezifischer Modelle effizient durchführbar macht. Die max. 25 Teilnehmer*innen benötigen einen Laptop und Internetzugang. Die Verwendung einer Maus wird empfohlen.

Forschungsinteressen der Beitragenden

Maximilian Wehner ist Wissenschaftlicher Mitarbeiter am Lehrstuhl für Künstliche Intelligenz und Angewandte Informatik sowie am Zentrum für Philologie und Digitalität „Kallimachos“ der Julius-Maximilians-Universität Würzburg. Forschungsinteressen sind die Literatur der Frühen Neuzeit, die OCR früher Drucke sowie die Entwicklung entsprechender Vermittlungskonzepte.

Dr. Michael Dahnke arbeitet als Wissenschaftlicher Mitarbeiter am Zentrum für Informations- und Medientechnologie der Universität Siegen. Seine Forschungsschwerpunkte bewegen sich in den Bereichen digitaler Editionsphilologie, Datenmodellierung im Rahmen von TEI sowie der OCR und der Modellierung gewonnener Textdaten.

Florian Landes ist als Wissenschaftlicher Mitarbeiter bei der Bayerischen Akademie der Wissenschaften beschäftigt. Seine Forschungsinteressen liegen in den Bereichen der OCR sowie der digitalen Rekonstruktion.

Robert Nasarek ist Wissenschaftlicher Mitarbeiter am Lehrstuhl für Wirtschafts- und Sozialgeschichte der Martin-Luther-Universität Halle-Wittenberg sowie des Zentrums für Wissenschaftsforschung der Nationalen Akademie der Wissenschaften Leopoldina. Seine Arbeit bewegt sich im Bereich der Wirtschafts- und Sozialgeschichte, OCR und Digital Humanities.

Christian Reul ist Kommissarischer Leiter der Digitalisierungseinheit des Zentrums für Philologie und Digitalität „Kallimachos“ der Julius-Maximilians-Universität Würzburg. Seine Forschungsschwerpunkte sind die OCR auf historischem Material sowie die Entwicklung von OCR-Software.

Fußnoten

1. https://www.abbyy.com/de-de/finereader/
2. https://www.uni-wuerzburg.de/zpd/ocr4all
3. https://github.com/OCR4all/LAREX
4. https://github.com/tmbdev/ocropy
5. https://github.com/Calamari-OCR/calamari
6. https://www.primaresearch.org/tools/PAGELibraries

Bibliographie

Breuel, Thomas M. / Ul-Hasan, Adnan / Al-Azawi, Mayce Ali / Shafait, Faisal (2013): High-Performance OCR for Printed English and Fraktur Using LSTM Networks, in: 12th International Conference on Document Analysis and Recognition: 683-687.

Reul, Christian / Christ, Dennis / Hartelt, Alexander / Balbach, Nico / Wehner, Maximilian / Springmann, Uwe / Wick, Christoph / Grundig, Christine / Büttner, Andreas / Puppe, Frank (2019): OCR4all – An Open-Source Tool Providing a (Semi-)Automatic OCR Workflow for Historical Printings, in: *ArXiv Preprints (submitted to MDPI – Applied Sciences)* https://arxiv.org/abs/1909.04032.

Rydberg-Cox, Jeffrey A. (2009): Digitizing Latin Incunabula: Callenges, Methods, and Possibilities, in: *Digital Humanities Quarterly* 3, 1 http://digitalhumanities.org:8081/dhq/vol/3/1/000027/000027.html.

Springmann, Uwe / Lüdeling, Anke (2017): OCR of historical printings with an application to building diachronic corpora: A case study using the RIDGES herbal corpus, in: *Digital Humanities Quarterly* 11, 2 http://www.digitalhumanities.org/dhq/vol/11/2/000288/000288.html.