Vom gedruckten Werk zu elektronischem Volltext als Forschungsgrundlage Erstellung von Forschungsdaten mit OCR-Verfahren

Boenig, Matthias
boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften

Herrmann, Elisa
elisa.herrmann@hab.de
Herzog August Bibliothek Wolfenbüttel

Hartmann, Volker
volker.hartmann@kit.edu
Steinbuch Centre for Computing / KIT

EINLEITUNG

In den vergangenen 30 Jahren ist ein beträchtlicher Teil des in Deutschland gedruckten Materials aus der Zeit von 1500 bis ca. 1850 in mehreren, durch die Deutsche Forschungsgemeinschaft (DFG) geförderten Kampagnen in den Verzeichnissen der im deutschen Sprachraum erschienenen Drucke des 16.-18. Jahrhunderts (VD16, VD17, VD18) zunächst nachgewiesen und seit 2006 digitalisiert worden. Zusätzlich vorliegender Volltext wird mittlerweile auf breiter disziplinärer Front als Schlüssel zu einer ganzen Reihe von geistes- und kulturwissenschaftlichen Forschungsfragen gesehen und gilt zunehmend als elementare Voraussetzung für die Weiterentwicklung der transdisziplinär arbeitenden Digital Humanities. Deshalb werden bereits an verschiedenen Stellen OCR-Verfahren angewendet; viele dieser Unternehmungen haben allerdings noch sehr starken Projektcharakter. Die informationswissenschaftliche Auseinandersetzung mit OCR kann an der großen Zahl wissenschaftlicher Studien und Wettbewerbe ermessen werden, die Möglichkeiten zur Verbesserung der Textgenauigkeit sind in den letzten Jahrzehnten enorm gestiegen. Der Transfer der auf diesem Wege gewonnenen, oftmals sehr vielversprechenden Erkenntnisse in produktive Anwendungen ist jedoch häufig nicht gegeben: Es fehlt an leicht nachnutzbaren Anwendungen, die eine qualitativ hochwertige Massenvolltextdigitalisierung aller historischen Drucke aus dem Zeitraum des 16. bis 19. Jahrhundert ermöglichen.

Auf dem DFG-Workshop „Verfahren zur Verbesserung von OCR-Ergebnissen“ (Deutsche Forschungsgemeinschaft 2014) im März 2014 formulierten Expertinnen und Experten daher folgende Desiderate um die Weiterentwicklung von OCR-Verfahren zu ermöglichen. Es bestehe eine dringende Notwendigkeit für freien Zugang zu historischen Textkorpora und lexikalischen Ressourcen zum Training von vorhandener Software zur Texterkennung bestehe. Ebenso müssen Open-Source-OCR-Engines zur Verbesserung der Textgenauigkeit weiterentwickelt werden, wie auch Anwendungen für die Nachkorrektur der automatisch erstellten Texte. Daneben sollten Workflow, Standards und Verfahren der Langzeitarchivierung mit Blick auf zukünftige Anforderungen an den OCR-Prozess optimiert werden. Als zentrales Ergebnis dieses Workshops stand fest, dass eine koordinierte Fördermaßnahme der DFG notwendig ist. Die „Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)“, kurz OCR-D, begann im September 2015 und versucht seitdem einen Lückenschluss zwischen Forschung und Praxiseinsatz, indem für die Entwicklungsbedarfe Lösungen erarbeitet und der aktuelle Forschungsstand zur OCR mit den Anforderungen aus der Praxis zusammengebracht werden.

ARBEITEN IM PROJEKT OCR-D

Das Vorhaben hat zum Ziel, einerseits Verfahren zu beschreiben und Richtlinien zu erarbeiten, um einen optimalen Workflow sowie eine möglichst weitreichende Standardisierung von OCR-bezogenen Prozessen und Metadaten zu erzielen, andererseits die vollständige Transformation des schriftlichen deutschen Kulturerbes in digitale Forschungsdaten in (xml-strukturierter Volltext) konzeptionell vorzubereiten. Am Ende des Gesamtvorhabens (d.h. unter Einschluss der Modulprojektphase) sollte ein in allen Aspekten konsolidiertes Verfahren zur OCR-Verarbeitung von Digitalisaten des schriftlichen deutschen Kulturerbes stehen und eine Dokumentation, die Antworten auf die damit verbundenen technischen, informationswissenschaftlichen und organisatorischen Probleme und Herausforderungen gibt sowie Rahmenbedingungen formuliert.

Das Projekt ist in zwei Phasen geteilt: In der ersten Phase hat das Koordinierungsgremium von OCR-D Bedarfe für die Weiterentwicklung von OCR-Technologien analysiert und sich intensiv mit den Möglichkeiten und Grenzen der Verfahren zur Text- und Strukturerkennung auseinandergesetzt. Zahlreiche Gespräche mit ExpertInnen aus Forschungseinrichtungen und Bibliotheken sowie Sichtung vorhandener Werkzeuge aber auch Betrachtung vorhandener Textsammlungen sowie aktueller und geplanter Digitalisierungsvorhaben

mündeten in der Erkenntnis, dass der Lückenschluss zwischen Wissenschaft und Praxis das primäre Desiderat im Bereich der Textdigitalisierung darstellt. Zudem hat sich im Lauf der ersten Projektphase eine technologische Wende auf dem Gebiet der Zeichenerkennung vollzogen - an die Stelle traditioneller Verfahren der Mustererkennung, die auf einer Segmentierung von Textabschnitten in Zeilen, Wörter und schließlich einzelne Glyphen basieren, die anschließend aufgrund charakteristischer Merkmale (z.B. Steigung an Kanten) erkannt werden, ist eine zeilenorientierte Sequenzklassifizierung auf Basis statistischer Modelle, insbesondere verschiedener Arten neuronaler Netze (sog. *Deep Learning*), getreten. Grund für diesen Technologiewechsel ist die vielfach nachgewiesene Überlegenheit segmentierungsfreier Erkennungsverfahren bezüglich der resultierenden Textgenauigkeit. Diese Überlegenheit gilt insbesondere für schwierige, historische Vorlagen. Dieser Technologiewandel hat sich bisher nicht oder nur äußerst begrenzt auf die Digitalisierungspraxis ausgewirkt. Der Grund dafür liegt vor allem in den bisher bestehenden Hürden beim Einsatz verfügbarer OCR-Lösungen auf Basis neuronaler Netze. Ohne weitreichende projektspezifische Anpassungen ist ein produktiver Einsatz derzeit nicht möglich. Das betrifft unter anderem die Erstellung passender Erkennungsmodelle, die durch das Trainieren eines neuronalen Netzes auf Basis ausgewählter Ground-Truth-Daten generiert werden. Dafür sind zum einen hochqualitativer und umfangreicher Ground Truth aber auch Erfahrungen bzgl. freier Parameter wie z.B. Anzahl der Trainingsschritte, Lernrate, Modelltiefe unabdingbar. Aus OCR-D heraus ist daher ein Datenset mit Trainings- und Ground-Truth-Daten entstanden, welches für Trainings und Qualitätsanalysen im Vorhaben selber genutzt wird aber auch durch andere Forschungsprojekte nachgenutzt werden kann. Neben der Qualität der Zeichenerkennung sind es vor allem Umfang und Korrektheit der strukturellen Annotationen, die die Utilität eines Volltexts für wissenschaftliche Kontexte determinieren. Auch im Bereich der automatischen Layouterkennung (OLR) gab es innerhalb des bisherigen Projektzeitraums vielversprechende Forschungsergebnisse durch den Einsatz innovativer statistischer Verfahren. Der Übertrag in die Praxis in Form nachnutzbarer Software ist hier jedoch noch nicht gegeben. Kommerzielle OCR-Lösungen ignorieren diesen Bereich weitestgehend und bieten nur minimale Strukturinformationen auf Seitenebene (Text, Tabelle, Abbildung etc.) an. Tiefergehende strukturelle Auszeichnungen (Kapitelstruktur, Bildunterschriften, Inhaltsverzeichnisse) werden daher manuell erfasst und in METS/MODS repräsentiert. Eine Verknüpfung zwischen Struktur und Volltext findet, obwohl technisch möglich, in vielen Digitalisierungsvorhaben nicht statt. Für die philologische, editorische oder linguistische Wissenschaftspraxis bedeutet das eine massive Einschränkung die bspw. eine sinnvolle Transformation in hochstrukturierte Formate wie TEI verhindert.

Die Erkenntnisse dieser Bedarfsanalyse mündeten in einem OCR-D-Funktionsmodell, welches den Rahmen für die Modulprojekt-Ausschreibung der DFG im März 2017 bot. Vor diesem Hintergrund wurden acht Modulprojekte bewilligt die seit 2018 an Lösungen zur Bildvorverarbeitung, Layouterkennung, Textoptimierung (inkl. Nachkorrektur), zum Modelltraining und zur Langzeitarchivierung der OCR-Daten arbeitet. Die Entwicklungen schöpfen dabei das Potential innovativer Methoden für den gesamten Bereich der automatischen Texterkennung für die Massenvolltextdigitalisierung von historischen Drucken aus. Sie werden anschließend nahtlos in den OCR-D-Workflow zur optimierten OCR-basierten Texterfassung integriert. Das so entstehende OCR-D-Softwarepaket steht damit Kultureinrichtungen wie Forschenden für die automatische Texterkennung als Open-Source-Software zur Verfügung.

Die meisten Arbeiten werden im Sommer 2019 abgeschlossen sein, aber bereits Anfang des Jahres wird die Alpha-Version einen Einblick in die zu erwartende Gesamtlösung bieten können.

ZIEL DES WORKSHOPS

Der Workshop soll neben der Vorstellung des Projektes und der Software die Gelegenheit bieten selber die Software zu testen und zugleich über Optimierungen und Anforderungen seitens der Wissenschaft an diese Technologien zu diskutieren. Teilnehmende erhalten somit einen exklusiven Einblick in die Entwicklungsarbeit und haben die Möglichkeit proaktiv auf die Arbeiten Einfluss zu nehmen, die Ihren späteren Forschungsalltag begleiten und verbessern soll.

PROGRAMM

Der Workshop gliedert sich in drei Abschnitte:

- Vorstellung des Projekts OCR-D, des Ground-Truth-Datensets und der Guidelines (30min)
- Demonstration der Eigenentwicklung und eines Test-Workflows (120min)
- Diskussion zu Anforderungen und Optimierungen aus Sicht der Digital Humanities (30min)

Der erste Abschnitt stellt die Hintergründe zum Vorhaben vor und geht auf Besonderheiten der Volltextdigitalisierung von historischen Beständen ein. Anschließend wird das Trainings- und Ground-Truth-Datenset präsentiert, das im Rahmen von OCR-D auf- und weiter ausgebaut wird. Besonders die dazu entwickelten Guidelines geben Hinweise für eine spätere Nachnutzung und die Erstellung eigener Ground-Truth-Daten in anderen Projekten. Der Fokus des Workshops liegt auf dem zweiten Abschnitt, in welche der derzeitige Entwicklungsstand präsentiert wird. Die benötigten Test-Dateien werden auf GitHub[1]

veröffentlicht. Abgerundet wird der Workshop durch eine Diskussionsrunde zu Anforderungen aus der Wissenschaft heraus an OCR-Techniken und die dafür eingesetzte Software.

VORAUSSETZUNG

Teilnehmende benötigen einen eigenen Laptop mit Internetanbindung und Ubuntu 18.04 als Betriebssystem. Alternativ kann auch Windows/Mac OSX mit der Software VirtualBox verwendet werden. Die VM wird den Teilnehmenden vom OCR-D-Projekt vor Ort zur Verfügung gestellt. Die Anzahl der Teilnehmenden ist auf 20-25 begrenzt. Python- und Linux-Kommandozeilen-Kenntnisse sind wünschenswert

Fußnoten

1. OCR-D Git-Hub: https://github.com/OCR-D/

Bibliographie

Deutsche Forschungsgemeinschaft (2014): Workshop *Verfahren zur Verbesserung von OCR-Ergebnissen*. Protokoll zu den Ergebnissen und Empfehlungen des Workshops. http://www.dfg.de/download/pdf/foerderung/programme/lis/140522_ergebnisprotokoll_ocr_workshop.pdf [Zuletzt abgerufen 07.01.2019]