Absichtserklärung

im Rahmen der Ausschreibung „*Implementierung der OCR-D-Software zur Volltextdigitalisierung historischer Drucke*“

Eine Ausschreibung im Rahmen des Förderprogramms „*e-Research-Technologien“*

Projekttitel „*Workflow für werkspezifisches Training auf Basis generischer Modelle mit OCR-D sowie Ground-Truth-Aufwertung“*

Antragsteller

Antragstellerin ist die Universitätsbibliothek Mannheim, vertreten durch die Ltd. Bibliotheksdirektorin Dr. Sabine Gehrlein.

Verantwortlicher Projektleiter ist Stefan Weil, kommissarischer Leiter der Abteilung Digitale Bibliotheksdienste.

Projektziele

Ausgangslage

Es besteht weitgehender Konsens, dass für den erfolgreichen Einsatz von OCR-D die Integration in bestehende Digitalisierungsworkflows von zentraler Bedeutung ist und dass dabei für die Layouterkennung und die Qualität der Texterkennung noch deutliches Verbesserungspotential besteht.

Die UB Mannheim möchte im OCR-D-Workflow die Texterkennung durch generische Modelle für Tesseract verbessern und dabei werkspezifisches Training für alle Anwender ermöglichen. Während für generische Modelle die Zeichenerkennungsraten typischerweise zwischen 90 und 95 % liegen, lassen sich durch werkspezifisches Training Zeichenerkennungsraten von über 98 % erreichen. Zudem können auch domänenspezifische Glyphen (beispielsweise mathematische oder astronomische Symbole[1] wie □ oder ⚷) nachtrainiert werden.

Training generischer Modelle

Für die Texterkennung mit künstlichen neuronalen Netzen – dem Stand der Technik für OCR – benötigt man Modelle, also Dateien mit den Parametern des neuronalen Netzes, die zuvor durch ein Training bestimmt wurden.

In der Vergangenheit ging man davon aus, dass für unterschiedliche Sprachen und Schriften jeweils spezifische Modelle notwendig sind. Das erwies sich immer wieder als problematisch, weil man für die OCR-Generierung die Sprache bzw. Sprachen des Textes angeben musste und zwischen Antiqua- und Frakturschrift zu wählen war, was oft zu Fehlern führte, denn Sprache und Schriftart können innerhalb eines Werkes wechseln. So findet man in den Alten Drucken der UB Mannheim häufig lateinische, griechische und französische Abschnitte in deutschsprachigen Büchern. Die Zeitung *Deutscher Reichsanzeiger* verwendet meist Frakturschrift, enthält aber auch englische, französische, spanische und italienische Texte, die ebenso wie manche Tabellen in Antiquaschrift gedruckt sind.

Gute generische Modelle für Texterkennung mit neuronalen Netzen benötigen keine Angabe der Sprache, da sie einerseits auch ohne Wörterbuch hohe Erkennungsraten erzielen und andererseits alle wesentlichen Zeichen für unterschiedliche Sprachen abdecken können.

Generische Modelle sind geeignet für einen breiten Einsatz mit unterschiedlichsten Werken, bei denen keine erhöhten Anforderungen an die Genauigkeit der OCR bestehen. Sie sind aber auch die Basis für werkspezifische Modelle, die daraus per Nachtraining (Finetuning) erzeugt werden.

Modelle, die auf Basis der GT4HistOCR-Daten[2] an der UB Mannheim trainiert wurden,[3] decken ein breites Schrift- und Zeichenspektrum für Drucke vom 15. bis 20. Jahrhundert ab. Sie haben gezeigt, dass es nicht unbedingt notwendig ist, spezifische Modelle für unterschiedliche Sprachen und Schriften

1 Vgl. https://de.wikipedia.org/wiki/Astronomisches_Symbol
2 Vgl. http://doi.org/10.5281/zenodo.1344132
3 Vgl. https://github.com/tesseract-ocr/tesstrain/wiki/GT4HistOCR

zu erzeugen, was mit zusätzlichem Aufwand verbunden wäre.

Im Projekt wird das optimale Training eines neuen generischen Tesseract-Modells, das bisherige GT4HistOCR-Modelle noch übertrifft, für das Werksspektrum von OCR-D angegangen. Dieses Modell kann frei nachgenutzt werden und ist leicht in andere Implementierungen von OCR-Workflows integrierbar.

Werkspezifisches Training

Unterstützt werden soll das werkspezifische Training von Modellen mit besonders hoher Erkennungsrate und domänenspezifischen Glyphen. Einrichtungen mit umfangreichen Werken mit einheitlichem Schriftbild können durch Transkribieren einer kleinen Anzahl von Seiten Daten bereitstellen, die in das Training eines spezialisierten Modells einfließen.

Ein automatisierter Arbeitsablauf soll nach manueller Transkription gemäß OCR-D-Transkriptionsrichtlinien und Bereitstellung der Transkriptionsergebnisse (Ground Truth) als PAGE XML oder Zeilenbildern und -texten anwenderfreundlich neue Tesseract-Modelle produzieren, die dann im OCR-D-Workflow für die Texterkennung verwendet werden.

Das Training der neuen Modelle erfolgt dabei wahlweise beim Anwender (Einrichtung, Dienstleister) oder auch optional für eine beschränkte Zahl von Anwendern und Nutzungsfällen auf Servern der UB Mannheim.

Für die Durchführung beim Anwender werden beispielsweise Erweiterungen von ocrd_all und Docker-Container bereitgestellt, die eine einfache Installation ermöglichen.

Für das werkspezifische Modelltraining bei der UB Mannheim bietet diese einen geeigneten Webdienst an, der für die vom Anwender bereitgestellte Ground Truth ein neues Modell erzeugt. Die UB Mannheim würde diese Daten auch zur weiteren Verbesserung des generischen Modelles nachnutzen, die Daten und neuen Modelle beispielsweise auf GitHub veröffentlichen und so ermöglichen, dass auch andere OCR-Software sie verwenden kann.

Details der Einbindung in den OCR-D-Workflow werden mit den anderen Projekten abgestimmt.

Korrektur und Aufwertung von Ground Truth

Beim Training von Tesseract-Modellen hat sich gezeigt, dass Fehler in den Ground-Truth-Daten mit erlernt werden und zu Fehlern in den OCR-Ergebnissen führen. Ebenso war eine wichtige Erkenntnis, dass praktisch jeder bisher untersuchte Ground-Truth-Datensatz eine gewisse Fehlerrate aufweist, also falsch abgeschriebene, fehlende oder überzählige Worte. Insbesondere mit Transkribus erzeugte Daten enthalten teilweise viele Fehler, weil sie oft nicht händisch transkribiert sind, sondern dabei OCR-Ergebnisse von ABBYY FineReader unzureichend manuell nachbearbeitet wurden. Auch entsprechen die Transkribus-Datensätze fast immer nur dem Level 1 gemäß der für OCR-D festgelegten Richtlinien zur Transkription von Ground Truth,[4] ignorieren also spezifische drucktechnische Aspekte und typographische Besonderheiten. Es ist aber möglich, mit Hilfe von hochwertiger OCR Fehler in Ground Truth zu finden und zu korrigieren. Auch die Aufwertung auf Level 2 mit Berücksichtigung schriftspezifischer Zeichen – beispielsweise durch Einführung des langen s „f“ – ist machbar. Beides wird schon jetzt durch das an der UB Mannheim entwickelte Tool GTCheck[5] unterstützt und soll im Rahmen des Projektes weiter ausgebaut werden.

Anwendungsszenario

Ein typisches Anwendungsszenario sieht wie folgt aus:

Für ein bestimmtes Werk soll Texterkennung mit erhöhten Anforderungen an die Qualität (Zeichen- und Worterkennungsraten) durchgeführt werden. Ein alternativer Anwendungsfall wäre ein Werk mit domänenspezifischen Glyphen, die von den vorhandenen Modellen nicht erkannt werden, deren Erkennung aber wünschenswert ist.

Hier müsste der Auftraggeber zunächst Transkriptionen repräsentativer Seiten des Werkes erstellen (lassen). Dafür gibt es bereits geeignete Werkzeuge, insbesondere Aletheia, LAREX und Transkribus. Die Ausgabe der Ground Truth erfolgt dabei typischerweise im Format PAGE XML.

4 Vgl. https://ocr-d.de/de/gt-guidelines/trans/transkription.html
5 Vgl. https://github.com/UB-Mannheim/GTCheck

Im nächsten Schritt ist diese Ground Truth mit vom Projekt bereitgestellten Werkzeugen zu prüfen. Erkannte Mängel (Fehler beim Abschreiben, neue Zeichen mit sehr geringer Häufigkeit) können dabei nachgebessert werden.

Qualifizierte Ground Truth fließt anschließend wieder als PAGE XML in den jeweiligen OCR-D-Workflow ein und wird dort verwendet, um ein werkspezifisches Tesseract-Modell zu trainieren. Dieses neu trainierte optimierte Modell wird dann für die Texterkennung verwendet.

Das gleiche Vorgehen ist auch für alle homogen gestalteten Werke einer Reihe anwendbar.

Vorarbeiten

Die UB Mannheim hat seit 2014 umfangreiche Erfahrungen im Einsatz und der Entwicklung von unterschiedlicher OCR-Software gesammelt und stellt auf GitHub mehrere Eigenentwicklungen aus dem OCR-Umfeld bereit.[6]

Im OCR-D-Modulprojekt „Tesseract als Komponente im OCR-D-Workflow“[7] (2018–2019) konnte die UB Mannheim die Software Tesseract durch zahlreiche Verbesserungen für den breiten Einsatz massentauglich machen. Mit den auf Basis von GT4HistOCR neu trainierten Tesseract-Modellen wurde erstmals eine Erkennungsqualität erreicht, die besser ist als mit kommerzieller Software.

Als Pilotbibliothek setzte die UB Mannheim OCR-D für die Volltexterkennung eigener Digitalisate ein. Mit dem in diesem Kontext entwickelten ocrd_all hat die UB Mannheim alle OCR-D-Komponenten unter einem Dach zusammengefasst und die Installation von OCR-D so wesentlich erleichtert. Auch nach dem Abschluss der Modulprojekt- und Pilotphase von OCR-D begleitet die UB Mannheim das Projekt weiter aktiv mit eigenen Beiträgen.

Neue Tesseract-Modelle hat die UB Mannheim inzwischen mit diversen Ground-Truth-Datensätzen wie GT4HistOCR, Austrian Newspapers, Neue Zürcher Zeitung, Fibeln des Georg-Eckert-Instituts, Jacob Grimms Weisthümer durchgeführt. Dabei wurde vorhandene Ground Truth nachkorrigiert und aufgewertet. Die entsprechende Dokumentation steht auf GitHub,[8] ebenso die zugehörigen Daten, soweit dies rechtlich möglich ist.

Gemeinsam mit der UB Tübingen berät die UB Mannheim Archive und Bibliotheken in Baden-Württemberg im Rahmen des Landesprojektes OCR-BW rund um das Thema Volltexterkennung auch beim Einsatz von OCR-D.[9]

Grobe Abschätzung des Mittelbedarfs

Für die Realisierung wird mit einem Gesamtaufwand von 36 Personenmonaten in der Personalkostenkategorie Doktorandin/Doktorand (TV/L E13) und 24 Personenmonaten für studentische Hilfskräfte über 24 Monate Laufzeit sowie Sachmitteln in Höhe von ca. 6.000 € für Reise- und Publikationskosten gerechnet.

Mannheim, 20.05.2020

Dr. Sabine Gehrlein
Universitätsbibliothek Mannheim

6 Vgl. https://github.com/UB-Mannheim/
7 Vgl. https://ocr-d.de/de/module-projects#optimierter-einsatz-von-ocr-verfahren--tesseract-als-komponente-im-ocr-d-workflow
8 Vgl. https://github.com/tesseract-ocr/tesstrain/wiki
9 Vgl. https://ocr-bw.bib.uni-mannheim.de/