

VERGLEICH DES UMGANGS MIT TRAININGSDATEN
UND MODELLEN IM DIGITALISIERUNGSPROZESS
VON HISTORISCHEN SCHRIFTEN

LEISTUNGSNACHWEIS 1

VON

ANDRÉS BAUMELER

TELEFON: 076 443 04 71, E-MAIL: ANDRES@BAUMELER.DEV

ALTE RIEDIKERSTRASSE 5C, 8610 USTER

BETREUER

CLEMENS NEUDECKER



Universität
Zürich UZH



UNIVERSITÄT ZÜRICH, PHILOSOPHISCHE FAKULTÄT /
ZENTRALBIBLIOTHEK ZÜRICH

CAS DATENMANAGEMENT UND INFORMATIONSTECHNOLOGIEN

Inhaltsverzeichnis

1	Einleitung	2
2	Grundlagen	4
2.1	Volltexttransformation	4
2.2	Einsatz von neuronalen Netzwerken für OCR	5
2.3	Trainieren von neuronalen Netzwerken	5
2.4	Handhabung Trainingsdaten und Modelle	6
3	Vergleich des Umgangs mit Trainingsdaten und Modellen	7
3.1	OCR-D	7
3.2	Transkribus	8
3.3	Umgang mit Trainingsdaten	9
3.4	Umgang mit Modellen	9
4	Schluss	10
	Abkürzungen	12
	Literaturverzeichnis	13

Kapitel 1

Einleitung

Digitalisierung von historischen Schriften ist ein umfassender Prozess welcher aus vielen Schritten besteht. Optical Character Recognition (OCR) nimmt dabei eine wichtige Rolle ein. OCR wird im Digitalisierungsprozess dazu verwendet Text von digitalisierten Dokumenten maschinenlesbar zu machen. Dadurch können die Dokumente im Volltext durchsucht werden und der Text für weitere Analysen verwendet werden. In den letzten Jahren konnte die Qualität der OCR Resultate durch den Einsatz von Maschinelernen, insbesondere neuronale Netzwerke, verbessert werden. So ist es heute möglich auch handgeschriebenen Text automatisch zu erkennen. Dies ermöglicht die Retrodigitalisierung von historischen Schriften in einem vorher nie dagewesenen Volumen.

Der Einsatz von Neuronale Netzwerke setzt voraus, dass entsprechende Trainingsdaten in genügender Qualität und Quantität verfügbar sind um die eingesetzten neuronalen Netzwerke zu trainieren. Das Ergebnis des Trainingsvorgangs wird in Modellen gespeichert. Diese Modelle können für verschiedene Szenarien wiederverwendet werden und ersparen ein erneutes Trainieren des neuronalen Netzwerks. Für historische Schriften haben sich dabei zwei Lösungen herauskristallisiert: OCR-D und Transkribus. In diesem Text soll anhand dieser zwei Lösungen aufgezeigt werden wie heute mit Trainingsdaten und Modellen im Bereich der Digitalisierung von historischen Schriften umgegangen wird. OCR-D setzt auf einen open Source und verteilten An-

satz während Transkribus auf einen closed Source und zentralisierten Ansatz setzt.

Im Ersten Kapitel wird aufgezeigt wie die Volltexttransformation abläuft und wie dabei neuronale Netzwerke für die Texterkennung eingesetzt werden. Weiter wird erläutert wie Neuronale Netzwerke trainiert werden und welche Rolle dabei die verwendeten Trainingsdaten spielen. Im Zweiten Kapitel werden die Frameworks OCR-D und Transkribus vorgestellt. Dabei wird aufgezeigt wie die beiden Frameworks mit Trainingsdaten und trainierten Modellen umgehen. In diesem Abschnitt wird auch auf die Vor- und Nachteile der beiden Lösungsansätze eingegangen. Es handelt sich hierbei nicht um einen wissenschaftlichen Vergleich sondern um einen Erfahrungsbericht aus der Anwendung der zwei Lösungen im Privatbereich.

Kapitel 2

Grundlagen

2.1 Volltexttransformation

Der Prozess der Volltexttransformation besteht aus mehreren Schritten wovon die eigentliche Texterkennung nur einer ist. Je nach Ausgangslage sind dabei unterschiedliche Schritte notwendig. Bei historischen Dokumenten insbesondere fallen bei der Vorbereitung verschiedene Schritte an. Ein Beispielhafter Prozess zur Volltexttransformation beinhaltet die Schritte:

- Seitentrennung
- Aufbereiten der einzelnen Seitentrennung
- Optische Layout Erkennung und Segmentierung der Seiten
- Aufbereiten der einzelnen Segmente
- Unterteilen der einzelnen Segmente in Textzeilen
- Aufbereiten der einzelnen Textzeilen
- Texterkennung auf den einzelnen Textzeilen
- Zusammenfügen der erkannten Texte
- Klassifizierung der erkannten Regionen
- Dokumentanalyse und Export

- Nachbearbeitung und ggf. manuelle Kontrolle und Korrektur der Resultate

Die Unterteilung in einzelne Textzeilen ist notwendig, da die eingesetzten Neuronale Netzwerke (NN) die Texterkennung auf Zeilenebene durchführen. Für die meisten dieser Schritte können bzw. werden NN eingesetzt. Für den Vergleich im Umgang mit Trainingsdaten wird aber auf den Schritt der Texterkennung fokussiert. Das Training von Modellen zur Segmentierung von Seiten oder zur Klassifizierung von Textblöcken wird hier nicht behandelt.

2.2 Einsatz von neuronalen Netzwerken für OCR

NN sind ein Subset des Bereichs Maschinelles Lernen (ML). Neuronale Netze werden oft als künstliche Modelle des menschlichen Gehirns betrachtet und verwenden eine Vielzahl von Algorithmen und Methoden, um Muster in Daten zu erkennen und Probleme zu lösen.

In den letzten Jahren hat die Entwicklung von neuronalen Netzen enorme Fortschritte gemacht, insbesondere durch den Einsatz von Deep Learning, einer Variante des neuronalen Netzes, die mehrere Schichten von Netzwerken verwendet, um komplexe Aufgaben auszuführen. Neuronale Netzwerke bestehen aus einer Eingangs- und Ausgangsschicht. Dazwischen gibt es eine oder mehrere sog. versteckte Schichten. In jeder Schicht gibt es Knoten (künstliche Neuronen) welche mit einem oder mehreren Knoten aus anderen Schichten verbunden sind. [8].

2.3 Trainieren von neuronalen Netzwerken

Begriffe Trainingsdaten, Ground Truth und Modelle erklären. Aufzeigen wie neuronale Netzwerke trainiert werden und wie die verwendeten Trainingsdaten das Endergebnis beeinflussen können. Aufzeigen wie Trainingsdaten und Modell zusammenhängen. Aufzeigen wie ein Modell aus Trainingsdaten entsteht.

2.4 Handhabung Trainingsdaten und Modelle

Aufzeigen wie Trainingsdaten und Modelle Strukturiert sind. Aufzeigen welche Methoden es heute gibt um Trainingsdaten und Modelle zu verwalten.

Kapitel 3

Vergleich des Umgangs mit Trainingsdaten und Modellen

3.1 OCR-D

OCR-D wird im Rahmen des DFG-Projekts OCR-D entwickelt und hat zum Ziel die Volltexttransformation Drucken aus dem Deutschen Sprachraum des 16. bis 18. Jahrhunderts konzeptionell und technisch vorzubereiten [7]. Das OCR-D Framework ist OpenSource. Das bedeutet der Source Code ist öffentlich einsehbar und kann von interessierten Personen auch modifiziert werden. OCR-D verwendet zur Verwaltung des Source Code ein Repository auf GitHub.[1]

OCR-D ist ein Framework welches mehrere Softwaremodule verbindet. Durch diese Module herangehendweise können die einzelnen Schritte im OCR Workflow durch unabhängige Softwarekomponenten abgedeckt werden. Diese Softwarekomponenten werden im OCR-D Framework als Prozessoren bezeichnet. Für den OCR Prozessor können die OCR Engines Tesseract, Ocropus, Kraken und Calamari eingesetzt werden. Der Export der OCR Resultate geschieht im ALTO Format.

Die Verwendung von OCR-D ist kostenfrei. Unterstützung beim Setup und der Anwendung kann aus der OCR-D Community bezogen werden.

Der Anwendungsbereich für OCR-D sind Institutionen welche in der Lage

sind die Lösung selbst zu Installieren und zu Konfigurieren. Die Installation kann dabei nativ oder via Docker Container erfolgen. Bei der nativen Installation wird das Programm vom Source Code gebaut und installiert. Dies geschieht mit den Werkzeugen make. Bei der Installation via Docker wird eine eine Maschine oder Server mit einer Docker Engine vorausgesetzt. Das Docker image wird dann via docker pull vom Repository geladen. Der Export der OCR Resultat ist als PDF oder im ALTO Format möglich.

3.2 Transkribus

Transkribus ist eine kommerzielle Plattform für Texterkennung, Transkription und das Durchsuchen von historischen Dokumenten. Transkribus wurde im Rahmen des Horizon 2020 EU-Projekts READ von einem Konsortium führender Forschungsgruppen aus ganz Europa unter der Leitung der Universität Innsbruck entwickelt. Die Plattform wird von der Genossenschaft READ-COOP betrieben und weiter entwickelt. [2]

Für die einzelnen Schritte Layout- und Strukturerkennung sowie Texterkennung setzt Transkribus auf Maschinelles Lernen. [3]

Für die Texterkennung muss bei Transkribus bezahlt werden. Die Bezahlung erfolgt mit Credits welche vorgängig über den Transkribus Shop gekauft werden müssen. [6]

Um Transkribus anzuwenden ist keine Installation notwendig. Dokumente können direkt über die Webseite von Transkribus unter <https://transkribus.ai> möglich. Dabei werden die Bilddaten an die Server von Transkribus übermittelt und durchlaufen dort die Schritte für die Volltexttransformation. Für kleinere Mengen an Seiten entfällt somit eine lokale Installation, es wird lediglich ein aktueller Browser vorausgesetzt.

Für grössere Volumen bietet Transkribus auch eine Client Anwendung an. Dieser sogenannte Expert Client ist eine Eclipse basierte Java Applikation welche lokal installiert wird. Über diesen Client lassen sich eine Vielzahl Parameter für den Prozess der Volltexttransformation einstellen. Der Client dient auch dazu Bilder zu Transkribus Servern hochzuladen. [5]

Für die Anbindung an andere System bietet Transkribus auch ein REST

API zum Hochladen von Dokumenten und zum konfigurieren der Volltexttransformation. [4].

Die Bearbeitung der Bilder erfolgt aber auch in diesem Verfahren auf den Servern von Transkribus.

3.3 Umgang mit Trainingsdaten

Aufzeigen wie die beiden Frameworks mit Trainingsdaten umgehen. Wie kann neue Ground hinzugefügt werden.

3.4 Umgang mit Modellen

Aufzeigen wie die beiden Frameworks mit Modellen umgehen. Wie können Modelle wieder verwendet werden. Wie können trainierte Modelle geteilt werden. Welche Daten von einem Modell sind einsehbar. Kann ein Modell weiter trainiert werden.

Bei OCR-D sind die Modelle welcher erstellt werden immer unter der Kontrolle der erstellenden Institution. Da es sich um Open Source Software handelt ist offen einsehbar wie ein Modell zustande kommt. Da OCR-D ein Framework ist welches verschiedenen Prozessoren für die Texterkennung einsetzt müssen Modelle für die einzelnen Modelle separat trainiert werden.

Die folgenden Informationen beziehen sich auf den Transkribus Expert Client in der Version 1.24.2. Transkribus unterscheidet bei Modellen zwischen Modellen für die Layouterkennungen und Modellen für die Texterkennung. Bei den Modellen für die Texterkennung können Modelle entweder für die CITLab Engine oder für die PyLaia Engine trainiert werden.

Kapitel 4

Schluss

OCR-D und Transkribus lösen ähnliche Probleme mit unterschiedlichen Ansätzen. OCR-D setzt auf Open Source während Transkribus eine mehrheitlich geschlossene Plattform ist.

Die Einstiegshürden für Transkribus sind gerade für Institutionen mit geringem Informatikverständnis um einiges tiefer als bei OCR-D. Bei Transkribus reicht eine Registrierung aus um erste Dokumente über den Browserclient zu digitalisieren. Ein initial Setup von OCR-D benötigt mehr technisches Verständnis, bietet dann aber mehr Konfigurations- und Anpassungsmöglichkeiten. Der Vorteil von OCR-D ist, dass sämtliche verwendeten Komponenten Open Source sind. Dadurch kann wird die Abhängigkeit von einer Institution verringert. Der Prozess kann zudem transparent dokumentiert werden indem die verwendeten Prozessoren inklusiver Versionen und Parameter notiert werden.

Bei der Handhabung von Trainingsdaten zeigen sich Unterschiede zwischen den zwei Frameworks. Bei OCR-D sind die trainierten Modelle offen zugänglich und können geteilt werden. Hier stellen sich Probleme der Kompatibilität. Modelle eines Prozessors sind nicht unbedingt mit einem anderen Prozessor nutzbar. Auch bei unterschiedlichen Prozessor- oder Engine Versionen kann es zu Inkompatibilitäten kommen. Bei Transkribus existiert zwar die Möglichkeit Modelle zu teilen. Der Zugriff auf die Rohdate eines Modells ist aber nicht möglich. Das verhindert den Einsatz eines Modells auf

einener anderen Plattform. Zugleich wird durch den Ansatz von Transkribus die Komplexität reduziert. Die Benutzenden brauchen sich keine Gedanken um die Kompatibilität der Modelle zu machen.

Beide Plattformen bieten einen guten Funktionsumfang und liefern beeindruckende Resultate in der Volltexttransformation. Sowohl der Open Source Ansatz von OCR-D als auch der Ansatz einer kommerziellen Plattform bei Transkribus machen Sinn. Beide Ansätze kommen mit ihren jeweiligen Vor- und Nachteilen. Es kann nicht abschliessend gesagt werden, welcher der zwei Ansätze besser ist. In der Evaluationsphase eines Digitalisierungsprojektes sollte deshalb sorgfältig anhand der Anforderungen des Projekts und der durchführenden Institution entschieden werden welche Lösung die geeignetere ist.

Abkürzungen

OCR	Optical Character Recognition
ML	Maschinenlernen
NN	Neuronales Netzwerk

Literaturverzeichnis

- [1] Ocr-d. Git Repository, 02 2023. URL <https://github.com/OCR-D>.
- [2] Wir sind read-coop. Website, 2023. URL <https://readcoop.eu/de/ueber-uns/>.
- [3] Transkribus. Website, 02 2023. URL <https://readcoop.eu/de/transkribus/>.
- [4] Rest-api. Website, 02 2023. URL <https://readcoop.eu/de/transkribus/docu/rest-api/>.
- [5] Transkribus herunterladen. Website, 02 2023. URL <https://readcoop.eu/de/transkribus/download/>.
- [6] Credits und preise. Website, 02 2023. URL <https://readcoop.eu/de/transkribus/credits/>.
- [7] Elisabeth Engl. Ocr-d kompakt: Ergebnisse und stand der forschung in der förderinitiative. *Bibliothek Forschung und Praxis*, 44(2):218–230, 2020. doi: doi:10.1515/bfp-2020-0024. URL <https://doi.org/10.1515/bfp-2020-0024>.
- [8] IBM. What is a neural network? Website, 2023. URL <https://www.ibm.com/topics/neural-networks>.