Vergleich der Framworks OCR-D und Transkribus zur volltextransformation 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, Philosophische Fakultät /
Zentralbibliothek Zürich

CAS Datenmanagement und Informationstechnologien

Inhaltsverzeichnis

Kapitel 1

Einleitung

Die Volltexttransformation von historischen Schriften ist ein umfassender Prozess welcher aus mehreren Schritten besteht. Optical Character Recognition (OCR) nimmt dabei eine wichtige Rolle ein. OCR wird verwendet um 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 Verfahren aus dem Maschinenlernen, insbesondere neuronalen Netzwerken, verbessert werden. So ist es heute möglich handgeschriebenen mit Text automatisch zu erkennen. Dies ermöglicht die Retrodigitalisierung von historischen Schriften in einem vorher nie dagewesenen Volumen.

Für die Volltexterkennung von historische Schriften existieren mehrere Lösungen. Zwei verbreitete Lösungen sind das OCR-D Framework und die Transkribus Platform. Dieser Text versucht die zwei Lösungen mit einander zu Vergleichen um Gemeinsamkeiten und Unterschiede aufzuzeigen. OCR-D und Transkribus verfolgen unterschiedliche Ansätze: OCR-D setzt auf einen open Source und verteilten Ansatz während Transkribus auf einen closed Source und zentralisierten Ansatz setzt.

Im Ersten Kapitel wird zur Übersicht aufgezeigt wie der Prozess der Volltexttransformation in einer modernen Lösung abläuft. Im Zweiten Kapitel werden die Frameworks OCR-D und Transkribus vorgestellt und miteinander

verglichen. Ziel ist es eine Hilfestellung für Institutionen zu bieten welche sich für eine dieser zwei Lösungen entscheiden möchten. Es handelt sich hierbei nicht um einen wissenschaftlichen Vergleich sondern um einen Erfahrungsbericht aus der Anwendung der zwei Lösungen im Privatbereich.

Kapitel 2

Grundlagen

Der Prozess der Volltexttransformation besteht aus mehreren Schritten wovon die eigentliche Texterkennung nur einer ist. Je nach Ausgangslage- und Material sind dabei unterschiedliche Schritte notwendig. Ein Beispielhafter Prozess zur Volltextransformation 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 einzelen 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 Fortschritte moderner Lösungen für die Volltextransformation sind zu einem grossen Teil auf den Einsatz neuronaler Netzwerke zurückzuführen. Die meisten Schritte im Prozess der Volltexttransformation können von

neuronalen Netzwerken unterstützt werden. Aus diesem Grund wird vertieft darauf eingeangen wie beiden Frameworks mit neuronalen Netzwerken umgehen.

OCR-D und Transkribus setzen neuronale Netzwerke für die Segmentierung und die Texterkennung ein. Die effektiven Workflows in den beiden Frameworks hängen von jeweiligen Setup und Anwendungsfall ab und entsprechen deshalb nicht dem obenstehenden beispiel Workflow. Gerade OCR-D bietet für einzelne Schritte im Prozess mehrere sogenannte Prozessoren mit unterschiedlichen Verfahren an.[7], [5]

Der Einsatz von neuronalen netzwerken 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. So kann ein für eine bestimmte Schriftart trainiertes Modell wiederverwendet werden wenn eines Tages weitere Dokumente in der gleichen Schriftart transformiert werden sollen. Trainingsdaten sind eine Sammlung von Beispielen, die zum Training von künstlichen Intelligenz Modellen verwendet werden. Sie bestehen aus Eingabe- und Ausgabedaten. Bei der Texterkennung können Trainingsdaten beispielsweise aus Bilder von Textzeilen und dem auf dem Bild enthaltenen Text in maschinenlesbarer Form bestehen.

Ground Truth (zu Deutsch: Wahrheit oder Wirklichkeit) ist ein Begriff der in der künstlichen Intelligenz verwendet wird um die korrekten Ausgaben für eine Eingabe zu beschreiben. Ground Truth sind Trainingsdaten für welche die Eingangs - und Ausgangsdaten verifiziert wurden. Das bedeuetet zu einem Eingangswert ist der korrekte Ausgangswert fesgehalten. Diese Ground Truth muss meist durch mühsame manuelle Arbeit erstellt werden denn nur so kann sichergestellt werden, dass das Modell auf einer korrekten Grundlage trainiert wird.

Im Kontext der neuronalen Netzwerke bezeichnet ein Modell die Konfiguration eines Neuronales Netzwerk (NN). Zur Konfiguration gehören die Anzahl und Anordnung der Neuronen, die Verbindungen unter den Neuro-

nen sowie die Gewichtung dieser Verbindungen. Beim Training werden diese Werte stetig verändert um die Ausgabendaten des NN möglichst nahe an die Ausgabedaten aus der Ground Truth zu bringen.

Ein Netzwerk wird trainiert, indem es mit Trainingsdaten gefüttert wird um daraus Ausgangsdaten gemäss der aktuellen Konfiguration zu produzieren. Die Ausgabedaten werden mit den Eingabedaten verglichen und das Netzwerk wird angepasst. Diese Anpassungen erfolgen automatisiert druch ein Trainingsprogramm. Im Zuge dieser Anpassung können etwa Verbindungen zwischen den Neuronen gelöscht oder neu angelegt werden. Auch die Gewichtung der einzelnen Verbindungen kann angepasst werden. Ein Training geht über mehrere Durchläufe (sogenannte Epochen). Am Ende kann anhand mehrere Metriken verifiziert werden wie nahe das trainierte Modell an den gewünschte Resultaten liegt.

Die Qualität der Trainingsdaten hat einen grossen Einfluss auf die Leistung des Netzwerks. Wenn die Trainingsdaten nicht repräsentativ für das zu lösende Problem sind oder schlecht gelabelt wurden, kann das Netzwerk fehlerhaft trainiert werden und schlechte Vorhersagen treffen. Daher ist es wichtig, qualitativ hochwertige Trainingsdaten zu haben, um ein leistungsfähiges Modell zu trainieren.

Kapitel 3

Vergleich

3.1 Vorstellung 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 [10].

Das OCR-D Framework ist OpenSource. Das beudeutet der Source Code ist öffentlich einsehbar und kann von interssierten Personen auch modifiziert werden. OCR-D verwendet zur Verwaltung des Source Code ein Repository auf GitHub.[3]

OCR-D ist ein Framwork welches mehrere Softwaremodule verbindet. Durch diese Modulare herangehensweise können die einzelnen Schritte im Prozess der Volltexttransformation durch unabhängige Softwarekomponenten abgedeckt werden. Diese Softwarekomponenten werden im OCR-D Framework als Prozessoren bezeichnet. Für einen Prozessschritt existieren meist mehrere Prozessoren mit unterschiedlichen Eigenschaften und Funktionsweisen. 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. [2]

3.2 Vorstellung 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. [6]

Um Transkribus anzuwenden ist keine Installation notwendig. Dokumente können direkt über die Webseite von Transkribus unter https://transkribus.ai möglich. Dabei werden die Bildaten an die Server von Trankribus übermittelt und durchlaufen dort die Schritte für die Volltexttransformation. Es entfällt somit eine lokale Installation, es wird lediglich ein aktueller Browser vorausgesetzt. Der Export der OCR Resultate ist bei Transkribus im ALTO Format möglich. [7]

Für die Texterkennung muss bei Transkribus bezahlt werden. Die Bezahlung erfolgt mit Credits welche vorgängig über den Transkribus Shop gekauft werden müssen. [9]

Für die Anbindung an andere System beitet Transkribus auch ein REST API zum Hochladen von Dokumenten und zum konfigurieren der Volltextransformation. Die Bearbeitung der Bilder erfolgt aber auch in diesem Verfahren auf den Servern von Transkribus. [8].

3.3 Umgang mit Trainingsdaten und Modellen

Bei OCR-D liegt es in der Verantwortung der Anwender die Trainingsdaten zu verwalten. Bei OCR-D sind die erstellten Modelle immer unter der Kontrolle der erstellenden Institution. Trainingsdaten können über eine Website wie das OCR-D Ground Truth Repository veröffentlicht und geteilt werden. [4]

Transkribus bietet die Möglichkeit eigene Ground Truth zum Training hochzuladen. Dies ermöglicht es ein Modell für einen spezifischen Anwen-

dungsfall zu trainieren. Für die eigenen Trainingsdaten und Modelle können Zugriffsrechte vergeben werden. So ist es möglich ein Modell zwar öffentlich zugänglich zu machen, die zugrundeliegenden Trainingsdaten aber privat zu halten, etwa wenn die Trainingsdaten aus rechtlichen Gründen nicht veröffentlicht werden dürfen. Für die Modelle welche von Transkribus selbst veröffentlicht wurden, sind die Trainingsdaten nicht zugänglich. Ein Export des Modells, etwa zur Verwendung in einer eigenen Installation oder als Backup, ist aber nicht möglich. Die Rohdaten zu einem Modell sind nicht einsehbar. Transkribus zeigt auf der Seite zu einem Modell eine Übersicht mit Statistiken und Informationen, wie etwa der CER-Rate oder der anzahl trainierter Epochen, zu einem Modell. [1]

3.4 Vor- und Nachteile

Der Vorteil von OCR-D ist, dass sämtliche verwendeten Komponenten Open Source sind. Dadurch wird die Abhängigkeit von einem bestimmten Lieferanten verringert. Es besteht nicht die Gefahr, dass ein einzelner Hersteller entscheiden kann die Lösung nicht mehr anzubieten. Ein weitere Vorteil von OCR-D ist, dass dank des modularen Aufbaus und der open Source Komponenten auch eigene Prozessoren entwickelt oder bestehende Prozessoren angepasst werden können. Dadurch ist es möglich eine Lösung aufzubauen welche exakt an die eigenen Bedürfnisse angepasst ist. Die Nachteile von OCR-D sind die steilere Lernkurve und das komplexe Setup. Bevor mit der Volltextransformation gestartet werden kann, muss ein Workflow aufgebaut und konfiguriert werden. Die Interaktion und die Konfiguration erfolgt bei OCR-D ausschliesslich über die Kommandozeile. Dies kann für Benutzende mit wenig IT Erfahrung problematisch sein. Weitere Herausforderungen stellen sich beim Betrieb auf der eigenen Hardware. Benutzende müssen sich selber darum kümmern die korrekte Hardware zur Verfügung zu haben.

Der Vorteil von Transkribus ist die flache Lernkurve und der einfache Einstieg. Ein aktueller Browser und ein Konto bei Transkribus genügen um mit der Volltexttransformation zu starten. Damit ist die Lösung auch für Benutzende ohne IT Kenntnisse sehr gut zugänglich. Durch den Platformansatz

müssen sich Benutzende zudem keine Gedanken um die Konfiguration und Update von Software und Hardware machen. Da es sich bei Transkribus um eine kommerzielle Lösung handelt hat der Benutzende den Vorteil eines klaren Ansprechspartners welcher bei etwaigen Supportfällen unterstützen kann. Die Nachteile von Transkribus liegen auch im Platformansatz. Der Export der eigenen Daten ist nur in dem von Transkribus angedachten Umfang möglich. Auch die Kontrolle über den Prozess ist nur möglich wo Transkribus die entsprechenden Möglickeiten im Portal bietet. Weil alle Daten zentralisiert durch Transkribus verwaltet werden besteht die Möglichkeit eines "Vendor Lock-In". Das bedeuetet es kann unter Umständen sehr schwierig werden in Zukunft zu einen anderen Anbieter zu wechseln.

Kapitel 4

Schluss

4.1 Zielgruppen

Für Transkribus sind die Hürden für den Start aber tiefer als mit OCR-D. Bei Transkribus reicht eine Registierung aus um erste Dokumente über den Browserclient zu digitalisieren. Der Expert Client welcher mehr Einstellungsmöglichkeiten als die Browserversion bietet, kann unkompliziert als Java Applikation installiert werden. Der Expert Client ist für alle gängigen Betriebssysteme verfügbar. Ein initial Setup von OCR-D benötigt mehr technisches Verständnis, bietet dann aber mehr Konfigurations- und Anpassungsmöglichkeiten. Die Installation kann entweder klassisch als Python Anwendung oder mittels Docker als Container erfolgen. Durch den Einsatz von Container kann OCR-D ebenfalls auf allen gängigen Platformen eingesetzt werden. Aus diesen Gründen ist Transkribus eher für kleinere Institutionen mit wenig bis gar keinem IT Know-How geeignet während für Institutionen mit ausgeprägtem IT Know-How und eigener Infrastruktur OCR-D wohl die geeignetere Lösung ist. OCR-D und Transkribus lösen ähnliche Probleme mit unterschiedlichen Ansätzen. OCR-D setzt auf Open Source während Transkribus eine mehrheitlich geschlossene Plattform ist.

4.2 Fazit

Es zeigen sich grundlegende Unterschiede zwischen den beiden Frameworks. Der Umgang mit Trainingsdaten und Modellen wird ganz unterschiedlich gehandhabt.Beide Frameworks 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.

In diesem Vergelich wurden die Genauigkeit und Geschwindigkeit der Erkennung, die Hardwareanforderungen sowie die Kosten nicht verglichen. Ein vergleich dieser Eigenschaften lässt sich nur für einen klar definierten Anwendungsfall durchführen. Ein vergleich der Kosten für die Durchführung eines Volltexttransformationsprojektes könnte interssant sein. Es ist denkbar, dass OCR-D Kostenmmässig besser skaliert als Transkribus.

Es kann nicht abschliessend gesagt werden, welches der zwei Frameworks besser oder geeigneter für die Volltexttransformation von historischen Texten ist. In der Evaluationsphase eines solchen Projektes sollte deshalb sorgfälltig anhand der Anfoderungen des Projekts, den Fähigkeiten der durchführenden Institution sowie den zur Verfügung stehenden Mittel und Infrastruktur entschieden werden welche Lösung die geeignetere ist.

Abkürzungen

OCR	Optical Character Recognition
ML	Maschinenlernen
NN	Neuronales Netzwerk

Literaturverzeichnis

[1] Öffentliche ai-modelle in transkribus. Website, 2023 02. URL https://readcoop.eu/de/transkribus/oeffentliche-modelle/.

[2] Ocr-d faq. Website, 02 2023. URL https://ocr-d.de/en/faq.

[3] Ocr-d core. Git Repository, 02 2023. URL https://github.com/OCR-D/core.

[4] Ocr-d ground truth repository. Website, 02 2023. URL https://ocr-d-repo.scc.kit.edu/api/v1/metastore/bagit/.

[5] Ocr-d workflows. Website, 02 2023. URL https://ocr-d.de/en/workflows.

[6] Wir sind read-coop. Website, 2023. URL https://readcoop.eu/de/ueber-uns/.

[7] Transkribus. Website, 02 2023. URL https://readcoop.eu/de/transkribus/.

[8] Rest-api. Website, 02 2023. URL https://readcoop.eu/de/transkribus/docu/rest-api/.

[9] Credits und preise. Website, 02 2023. URL https://readcoop.eu/de/transkribus/credits/.

[10] 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.