Vergleich von 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

Ziel der Volltexttransformation von historischen Dokumenten ist es eine digitale Version des Textes, meist inklusive Layout Informationen, zu erstellen. Dadurch können die Dokumente im Volltext durchsucht werden und der Text für weitere Analysen verwendet werden. In den letzten Jahren konnte die Qualität der Resultate durch den Einsatz von Verfahren aus dem Maschinenlernen, insbesondere neuronalen Netzwerken, verbessert werden. So ist es heute möglich handgeschriebenen Text mit grosser Genauigkeit automatisch zu erkennen. Für die Volltexterkennung von historischen Schriften existieren mehrere Lösungen. Zwei verbreitete Lösungen sind das OCR-D Framework und die Transkribus Platform. Dieser Text versucht die zwei Lösungen miteinander zu Vergleichen um Gemeinsamkeiten und Unterschiede aufzuzeigen. OCR-D und Transkribus sind beides Lösungen zur Unterstützung des Volltexttransformationsprozesses mit Fokus auf historischen Schriften, verfolgen dabei aber unterschiedliche Ansätze: OCR-D setzt auf einen open Source und verteilten Ansatz während Transkribus auf einen mehrheitlich closed Source und zentralisierten Ansatz setzt. Der Vergleich findet auf einer technischen Ebene statt. Aspekte wie Kosten oder Performance werden nur am Rande behandelt.

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

In diesem Kapitel werden die Grundlagen der Volltexttransformation kurz beschrieben um im späteren Vergleich eine bessere Einordnung der Unterschiede zu ermöglichen. 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 meisten Schritte im Prozess der Volltexttransformation können von neuronalen Netzwerken unterstützt werden. Aus diesem Grund wird im Vergleich vertieft darauf eingeangen wie beiden Lösungen 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 zwingend nicht dem obenstehenden beispiel Workflow. Gerade OCR-D baietet 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. Bei der Texterkennung können Trainingsdaten beispielsweise aus Bilder von Textzeilen oder ganzen Buchseiten 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. Bei der Texterkennung kann Groundtruth etwa aus Bilder von Textzeilen mit dem darauf enthaltenen Text in digitaler Form bestehen.

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 durch 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 verschiedenen Metriken verifiziert werden wie nahe das trainierte Modell an den gewünschte Resultaten liegt.[12]

Für eine Lösung welche die Volltexttransformation unterstützt, ist es wichtig, dass Modelle gespeichert und wiederverwendet werden können. So kann über die Zeit von bereits verarbeitetem Material gelernt und die Erkennung verbessert werden. Ebenfalls wichtig ist der Umgang mit Trainingsdaten und Ground Truth. Gerade Ground Truth ist oftmals Arbeitsintensiv in der Herstellung. Eine Veröffentlichung von Ground Truth ist wünschenswert damit auch andere Parteien von der investierten Arbeit profitieren können.

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 [11].

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. Das OCR-D Frameworks stellt die Mittel zur Verfügung die einzelnen Komponenten untereinander zu verbinden und abzustimmen. Das OCR-D Framework ist Open Source. Teilweise werden Komponenten verwendet welche nicht von OCR-D selbst entwickelt wurden, es handel sich aber auch hierbei um Open Source Komponenten. Das beudeutet der Quellcode ist öffentlich einsehbar und kann von interssierten Personen auch modifiziert werden. OCR-D verwendet zur Verwaltung des Quellcode ein Repository auf GitHub.[1]

Die Verwendung von OCR-D ist kostenfrei, es fallen keine Lizenz- oder anderweitige Nutzungsgebühren an. Es muss allerdings bedacht werden, dass für den Betrieb Hardware notwendig ist welche vom Anwender selbst zur Verfügung gestellt werden muss. [3]

[1] https://github.com/OCR-D/core

Unterstützung beim Setup und der Anwendung kann aus der OCR-D Community bezogen werden. Nützliche Ressourcen sind dabei der OCR-D Gitter Kanal.[2]

Die einzelnen Softwarekomponenten werden im OCR-D Framework als Prozessoren bezeichnet. Innerhalb des OCR-D Framework existieren Prozessoren für sämtliche Schritte aus dem in Kapitel 2 dargestellten Prozess. Für einen Prozessschritt existieren meist mehrere Prozessoren mit unterschiedlichen Eigenschaften und Funktionsweisen. Für den Optical Character Recognition (OCR) Prozessor beispielsweise können die OCR Engines Tesseract, Ocropus, Kraken und Calamari eingesetzt werden.

Die Installation kann entweder klassisch als Python Anwendung oder mittels Docker als Container erfolgen. OCR-D ist unter Linux zu Hause, kann aber mit Technologien wie Container oder Virtualsierung auf allen gängigen Betriebssystemen betrieben werden. [1]

OCR-D nutzt für die Resultate der Layout und Texterkennung das PAGE-XML Format. Ein Export der Resultate ist auch in den Formaten ALTO, hOCR und ABBYY FineReader XML möglich. [3]

3.2 Vorstellung Transkribus

Transkribus ist eine 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]

Transkribus ist eine Lösung welche auf der Infrastruktur der READ-COOP Genossenschaft betrieben wird und gegen Bezahlung genutzt werden kann. Der Client für die Interaktion mit der Plaftorm ist Open Source. Der Kern der Platform ist aber nicht Open Source. Für die Texterkennung muss bei Transkribus bezahlt werden. Die Bezahlung erfolgt mit Credits welche

[2]https://gitter.im/OCR-D/Lobby

vorgängig über den Transkribus Shop gekauft werden müssen. [9]

Trankribus deckt alle Prozessschritte des im Kapitel 2 dargestellten Beispielprozesses ab. Darüber hinaus bietet die Platform Möglichkeiten zum Verwalten und Zusammenarbeiten von digitalen Dokumenten. [7]

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. Transkribus bietet die Möglichkeit eine Client Anwendung, den Expert Client, lokal zu installieren. Der Expert Client welcher mehr Einstellungsmöglichkeiten als die Browserversion bietet, wird als Java Applikation installiert und ist damit auf allen gängigen Betriebssystemen verfügbar. Für die Anbindung an andere System beitet Transkribus 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].

Transkribus bietet die Möglichkeit Dokumente auf einer 'read & search' Webseite zu veröffentlichen. Damit können Dokumente unkompliziert online zur Verfügung gestellt werden.[3] Ein Beispiel für eine solche 'read & search' Webseite ist die 'iurisprudentia' Edition des Lehrstuhl für Privatrecht, Schwerpunkt ZGB der Universität Zürich.[4] Unabhängig davon ist der Export der generierten Resultate als PDF oder im ALTO Format möglich. [7]

3.3 Umgang mit Trainingsdaten und Modellen

Wie in Kaptiel 2 beschrieben spielen Verfahren aus dem Maschinenlernen und damit Trainingsdaten und Modelle eine grosse Rolle für das Endresultat. Bei

[3] https://readcoop.eu/de/readsearch/
[4] https://rwi.app/iurisprudentia/de

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. Da OCR-D verschiedene Prozessoren mit verschiedenen Technologien einsetzt, sind die Trainingsdaten und Modelle nicht beliebig austauschbar. Es müssen immer die Anforderungen des jeweiligen Prozessors berücksichtigt werden. [4]

Transkribus bietet die Möglichkeit eigene Ground Truth zum Training hochzuladen. Dies ermöglicht es ein Modell für einen spezifischen Anwendungsfall zu trainieren. Ebenfalls bietet Transkribus die Möglichkeit ein bereits bestehendes Basismodell weiter für die eigenen Dokumente zu trainieren. Damit kann die Traininigszeit verkürzt werden. 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. Die Trainingsdaten für die von Transkribus trainierten Modelle sind teilweise öffentlich verfügbar. [10], [2]

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. [2]

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 Softwarelieferanten verringert. Ein weiterer 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 fast ausschliesslich über die Kommandozeile oder mit Konfigurationsfiles. 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 und Softwareabhängigkeiten 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. Die Verwaltung und Versionierung von Modellen wird durch die Platform ebenfalls vereinfacht. Durch die zentrale Ablage der Modelle können Modelle als Basismodelle genutz und zu einem späteren Zeitpunkt weiter trainiert werden. Da es sich bei Transkribus um eine kommerzielle Lösung handelt hat der Benutzende den Vorteil eines klaren Ansprechspartners welcher bei etwaigen Supportfällen unterstützen kann. Ein weitere Vorteil von Transkribus liegen in den Dienstleistungen welche über die eigentliche Volltexttransformation hinausgehen. Transkribus bietet die Möglichkeit hochgeladene Dokumente zu teilen und zu veröffentlichen. Dabei muss man sich als Benutzer nicht um das Bereit

Die Nachteile von Transkribus liegen wie die Vorteile ebenfalls 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 schwierig werden in Zukunft zu einem anderen Anbieter zu wechseln oder bereits erstellte Texte auf eine andere Platform zu verschieben.

Kapitel 4

Schluss

4.1 Zielgruppen

Wer die Zielgruppen der zwei Lösungen sind, hängt vom Blickwinkel ab. Nachfolgende werden die zwei Lösungen au dem Blickwinkel des notwendigen IT Know-How sowie der zu verarbeitenden Volumen betrachtet.

Für erste Erfolge mit Transkribus ist relativ wenig IT Know-How notwendig da die Infrastruktur und das Setup grösstenteils durch Transkribus kontrolliert sind. Bei Transkribus reicht eine Registierung aus um erste Dokumente über den Browserclient zu digitalisieren. Ein initial Setup von OCR-D benötigt mehr technisches Verständnis, da verschiedene Komponenten benötigt werden welche alle zuerst gemäss den eigenen Anforderungen konfiguriert werden müssen. Dafür bietet OCR-D dann mehr Möglichkeiten genau auf die eigenen Anforderungen einzugehen und den Prozess genau so zu gestalten wie dieser benötigt wird. Damit 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.

Die Zielgruppen hängen auch davon ab welche Volumen verarbeitet werden sollen. Für kleinere bis mittlere Volumen (~1-500 Seiten) ist Transkribus aufgrund des einfacherern Einstiegs und der flacheren Lernkurve im Vorteil. Für grössere Projekte ist OCR-D etwas im Vorteil, da die modulare und offe-

ne Architektur bessere möglichkeiten zur Automatisierung bietet. Trankribus bietet zwar ein REST API welches sich nutzen lässt um Dokumente automatisiert zu verarbeiten. Das setzt aber voraus, dass entsprechende Software vorhanden ist welche das REST API von Transkribus ansprechen kann. Bei OCR-D hingegen lassen sich die Komponenten über etablierte Mechanismen wie Shell Scripting und Pipelines untereinander und mit OCR-D fremden Lösungen integrieren.

4.2 Fazit

Es zeigen sich einige Gemeinsamkeiten aber auch grundlegende Unterschiede zwischen den beiden Lösungen. Beide Frameworks bieten einen guten Funktionsumfang und lieferen beeindruckende Resultate in der Volltexttransformation. Der Umgang mit Trainingsdaten und Modellen wird ganz unterschiedlich gehandhabt. Sowohl der Open Source Ansatz von OCR-D als auch der Ansatz einer kommerziellen Plattform bei Transkribus machen Sinn. Insbesondere die Dienstleistungen welche über die Volltexttransformation hinausgehen, können Transkribus für viele Institutionen attraktiv machen. Beide Ansätze kommen mit Vor- und Nachteilen welche je nach Situation und Ausgangslage unterschiedliche gewichtet werden müssen.

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 als weiterführende Fragestellung interssant sein. Es ist denkbar, dass OCR-D Kostenmmässig besser skaliert als Transkribus.

Es kann nicht abschliessend gesagt werden, welches der zwei Lösungen besser oder geeigneter für die Volltexttransformation von historischen Texten ist. Beide Lösungen sind sicherlich geeignet für den produktiven Einsatz und haben das in mehreren Projekten in der Praxis demonstriert. In der Evaluationsphase eines Volltexttransformationsprojektes sollte deshalb sorgfältig 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] Ocr-d setup. Website, 2023 02. URL https://ocr-d.de/en/setup.

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

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

[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] Markus Diem, Florian Kleber, Stefan Fiel, Tobias Grüning, and Basilis Gatos. ScriptNet: ICDAR 2017 Competition on Baseline Detection in Archival Documents (cBAD), January 2017. URL https://doi.org/10.5281/zenodo.1491441. This project has received funding

from the European Union's Horizon 2020 research and innovation programme under grant agreement No 674943.

[11] 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.

[12] IBM. What is a neural network? Website, 2023. URL https://www.ibm.com/topics/neural-networks.