Best-practices zur Erkennung alter Drucke und Handschriften. Die Nutzung von Transkribus large- und small-scale

Hodel, Tobias
tobias.hodel@uzh.ch
Universität Bern, Schweiz

In den vergangenen Jahren konnte die automatisierte Erkennung sowohl handschriftlicher als auch alter Druckschriften stark verbessert werden. Sowohl für Handschriften als auch für alte Drucke hat sich der Einsatz von Handwritten Text Recognition (HTR) bewährt, die auf dem Einsatz neuronaler Netze beruht.[1] Führend in der Implementierung der Technologie ist die Plattform Transkribus, die im Rahmen von Projekt READ zwischen 2016 und 2019 stetig weiter entwickelt wurde (Muehlberger u. a. 2019) und auf der mittlerweile (Stand Ende 2019) mehr als 1‘000‘000 Dokumentenseiten bearbeitet wurden.[2] Die Verbesserung der Technologie führte gleichzeitig zur Einführung von unterstützenden Tools und Methoden, die reine Texte entweder mit höherer Genauigkeit suchbar machen oder strukturierende Maßnahmen ermöglichen. Es spielt also eine Rolle, welche Resultate erreicht werden sollen und welche Ziele der vorgesehene Zugriff hat. Im Rahmen des Papers werden drei Herangehensweisen skizziert, die ausgehend von unterschiedlichen Zielvorstellungen andere Aufbereitungsschritte und die Allokation von Ressourcen an verschiedenen Stellen nötig machen. Das Paper fokussiert auf die Nutzung von Transkribus, da die Software frei nutzbar und aufgrund des GUI innerhalb von zwei Arbeitstagen ohne Vorkenntnisse erlernbar ist (siehe dazu auch die How-To Guides: READ 2019). Darin eingeschlossen ist die Anfertigung eigener Modelle zur Erkennung von Handschriften bzw. alter Drucke.

Verbesserung der Handschriftenerkennung

Handschriftenerkennung ist seit den 1990er Jahren und dem Aufkommen der OCR (Optical Character Recognition) ein Forschungsfeld der Computerwissenschaften. Nach einer frühen Euphorie folgte bis vor fünf Jahren eine Ernüchterung, da die erzielten Resultate, die häufig auf statistischen Modellen (insbesondere Hidden Markov Models) basierten, für den Einsatz in der Praxis ungenügend waren. Zeichenfehlerquoten von bestenfalls 16% CER (Character Error Rate) zeigten zwar die Chancen auf, der erkannte Text war aber weder lesbar noch sinnvoll durch Postkorrektion aufzubereiten (Sánchez et al. 2013). Erst der Einsatz neuronaler Netze (erst rekursive, später konvolutionale) führte dazu, dass die Fehlerquote auf unter 12% CER gedrückt werden konnte (Leifert et al. 2016). Ab der Schwelle um 12% wird die Korrektur von erkanntem Text gegenüber von händisch erstellten Transkriptionen ökonomisch sinnvoll. Gleichzeitig sind die Resultate ab 12% für Menschen insofern nützlich, da die Navigation im Text, insbesondere für Personen mit Kenntnissen der Dokumente, rasch und zielsicher möglich ist.

Zugriffsformen

Obwohl geisteswissenschaftliche Forschung häufig „Text“ im Fokus hat, ist die Diskussion, was darunter verstanden wird, bereits mehrfach in Bezug zu digitalen Editionsformen in den Digital Humanities breitgetreten worden (Sahle 2013, zu Texterkennung Hodel 2018). Aus der Perspektive von Forschenden, die Text als Grundlage nutzen, werden gleichzeitig unterschiedliche Fragen an das aufbereitete Material gestellt. Ob etwa ein Dokument nur durchsucht oder aber mit *text-mining* Methoden ausgewertet werden soll, führt zu unterschiedlichen Anforderungen an die Texterkennung. Zwischen den beiden Polen besteht auch die Möglichkeit nur visuell abgegrenzte Textteile auszuwerten. Für alle diese Zielvorstellungen unterscheidet sich der Aufwand für die Aufbereitung der Materialien.

Um eine Vergleichbarkeit herzustellen, lohnt sich eine systematisierte Sicht auf den Workflow mit Angaben, wo der Grossteil der Arbeit anfällt. Es werden daher im Folgenden Fragen nach Ablauf und Umfang der Arbeit sowie nach Schwierigkeiten/Probleme beschrieben.

Nicht thematisiert werden unterschiedliche Möglichkeiten bei Upload sowie Exportformate und insgesamt Fragen der Nachbereitung.

Hohe Textgüte als Ziel

Der klassische Zugriff zielt darauf ab, mit möglichst wenig Aufwand eine möglichst gute Texterkennung zu erzielen. Das Training von passgenauen Modellen steht dabei im Vordergrund. Aufbauend auf der Erkennung können *text-mining* Technologien ebenso eingesetzt werden, wie die Weiterverarbeitung als digitale Edition, etwa mit textkritischer Kommentierung oder durch die Annotation von *named entities*.

Ablauf

Primär muss möglichst viel Trainingsmaterial bereitgestellt werden, das bereits früh im Arbeitsprozess in

Modelle umgesetzt (= trainiert) wird. Generische Modelle spielen eine untergeordnete Rolle, da diese die Qualität der passgenauen Modelle nicht erreichen. Die Arbeitsweise ist iterativ, das heisst nach Aufbereitung von bereits 3'000 Wörtern lohnt sich die Herstellung eines Modells. Darauf aufbauend werden weitere Seiten erkannt und korrigiert. Der Prozess wird gestoppt, sobald die Verbesserung des Modells sich im Zehntelprozent Bereich bewegt (siehe dazu auch unten: Evaluation von trainierten Modellen).

Spezifische Typen von Layoutanalysen spielen keine Rolle.

Umfang der Arbeit

Gute Resultate (Zeichenfehler unter 5%) werden bei Dokumenten von derselben Hand mit 10'000 Wörter erreicht. Trainings können selbständig gestartet und überprüft werden.[3]

Schwierigkeiten/Probleme

Sobald unterschiedliche Hände in den Dokumenten erkannt werden sollen, ist eine Erhöhung der Trainingsmaterialien notwendig.

Semantische Textsegmentierung als Ziel

Dokumententypen können aufgrund von schematischem Aufbau nur in Teilen interessant sein, d.h. nur ein visuell abgesetzter Teil soll ausgewertet werden. Einzelne Textteile, bspw. Marginalien mit inhaltlichen Zusammenfassungen oder Fussnoten, werden für spezifische Forschungszwecke identifiziert.

Ablauf

Ein spezifischer Layouttyp wird trainiert. Danach wird unabhängig davon ein Textmodell entwickelt (analog zu „Hohe Textgüte“). Extraktion von Textregionen bzw. Einbindung in externe (webbasierte) Applikationen ist möglich.

Umfang der Arbeit

Mindestens 100, besser 200-300 Vorkommen der zu identifizierenden Teile (bspw. Textregionen). Zusätzlich müssen Aufwände für die Aufbereitung von Modellen veranschlagt werden.

Schwierigkeiten/Probleme

Das Training der semantischen Layouterkennung ist erst experimentell in Transkribus implementiert und schlecht dokumentiert (das Training kann auch extern über ein eigene Tool erfolgen [Ares Oliveira u. a. 2018; Quirós 2017]). Die Technologie ist insgesamt noch experimentell. Die Extraktion spezifischer Textregionen erfordert zudem Erfahrung im Umgang mit der REST API von Transkribus.

Suche in grossen Textbeständen als Ziel

Als regelmässige Anforderung wird die Suche in großen Dokumentenkorpora vorgegeben. Dazu scheint zwar ein probates Mittel die Erkennung mit hoher Textgüte zu sein, da für Handschriften und alte Drucke die Genauigkeit von OCR nicht gegeben ist, drängt sich ein alternativer Zugriff auf. Mittels Suche in allen vom Algorithmus erkannten Varianten, kann auch mit nicht passgenauen Modellen eine hohe Trefferquote (hoher *recall*) erreicht werden.

Ablauf

Primär wird ein möglichst passendes Model identifiziert. Einige generische Modelle (etwa für mittelalterliche Buchschriften oder lateinische Schriften aus den Niederlanden sowie deutsche Kurrent) sind bereits publiziert und in Zukunft werden noch weitere Modelle veröffentlicht. In einem zweiten Schritt werden einige wenige typische Seiten als Validierungsseiten aufbereitet bzw. sog. Samples (zufällig ausgewählte Zeilen, die eine statistisch valide Aussage ermöglichen) erstellt. Aufgrund der eruierten Zeichenfehlerquote in den Validierungssets, ist es möglich Aussagen über die erwartete Trefferquote zu machen. Ab Fehlerquoten von 15% CER und weniger, wird der Recall (Umfang aller gefundenen, möglichen Treffer) 99% betragen und somit werden nur wenige Treffer verpasst.

Umfang der Arbeit

Beschränkt sich vorwiegend auf die Identifikation passender Modelle und der Herstellung von Validierungssets (Validierungsseiten oder Samples bestehend aus einzelnen Zeilen) zur Validierung der Ergebnisse. Die Auswertung der Treffer mit Identifikation von *false-positive* Treffern am Ende des Prozesses bedarf manueller Arbeit.

Schwierigkeiten/Probleme

Die Auswertenden der Trefferliste müssen die Handschrift/den Druck selbst lesen können, um korrekte Treffer zu identifizieren. Die Suche erfolgt in den Erkenntabellen (sog. confidence matrices), da diese nicht

durch etablierte Datenbanksysteme etabliert werden, ist die Performanz niedrig und Suchen in mehr als 1'000 Seiten dauern mehrere Minuten (50'000 Seiten werden in knapp 3 Stunden durchsucht). Die Abfragen müssen in Transkribus erfolgen bzw. über die REST API.

Visualisierung

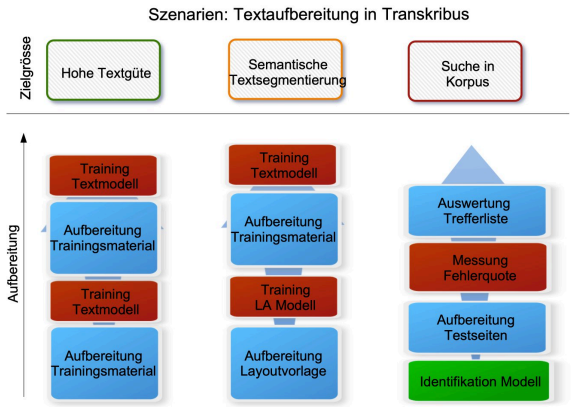

Abbildung 1: Visualisierung der Szenarien zur Textaufbereitung in Transkribus. Abbildung des Autors, CC-BY .

Trainingskurven: Evaluation von trainierten Modellen

Eine Aufgabe, die auch von Geisteswissenschaftlern mit nur bedingten technischen Vorkenntnissen übernommen werden kann, ist das Training von Textmodellen. Dazu muss ein Trainingsset (ca. 90% aller aufbereiteten Seiten) und ein Validierungsset (ca. 10% der Seiten) definiert werden. Das Training erfolgt danach auf den Servern in Innsbruck, es gibt nur zwei Optionen, die angepasst werden können. Erstens kann die Anzahl der Epochen (siehe unten) angepasst werden und zweitens können bereits vorhandene Modelle als Basismodelle gewählt werden.

Im Trainingsmodus erstellt ein Fehlertool Kurven, die anzeigen wie gut das Training ablief. Anhand dieser Trainingskurven lässt sich abschätzen, inwiefern ein Modell noch verbessert werden kann bzw. gar ein Re-training notwendig ist, da sich das Netz nicht wunschgemäss verbesserte.

Drei Begriffe müssen zum Verständnis vorgängig geklärt werden: **Neuronale Netze** stammen aus dem Bereich des maschinellen Lernens und versuchen aufgrund von Trainingsmaterial (Input und gewünschter Output) einen wertenden Algorithmus (ein Netz basierend auf je nach Input unterschiedlich reagierenden Speicherzellen) zu entwickeln (*trainieren*), der den gewünschten Ausgaben möglichst nahe kommt (Schöch 2017). **Epochen** meint die Anzahl an Wiederholungen, mit denen ein Netz mit denselben Trainingsdaten zwecks Verbesserung gefüttert wird. Am Ende jeder Epoche wird das Validierungsset durch das Netz erkannt und eruiert, welche Resultate erreicht worden wären. Dadurch entstehen zwei **Kurven** (in den Abbildungen rot für das Validierungsset und blau für das Trainingsset), die Aussagen über die Fähigkeit eines Netzes machen.

Trainings- und Validierungskurve divergiert stark

Wenn sich die Trainingskurve während dem gesamten Training stetig verbessert, das Validierungsset jedoch auf einer (weit) schlechteren Fehlerquote stehen bleibt, spricht man von *overfitting*. Das bedeutet, das Modell lernt die Trainingsseiten auswendig, ohne dass die Fähigkeit zur Erkennung der Zeichen wirklich eingelernt wird. Probates Mittel um den Effekt zu reduzieren, ist das Bereitstellen von mehr Trainingsmaterial.

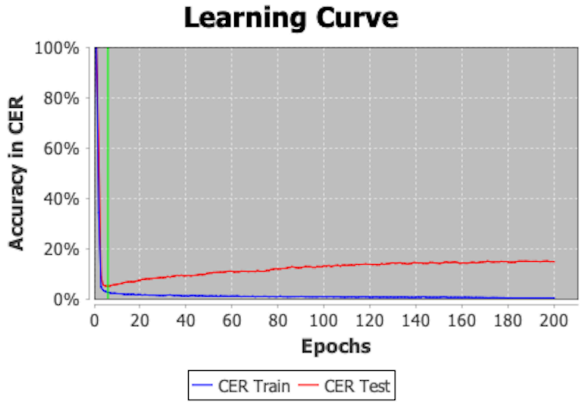

Abbildung 2: Lernkurve mit zuwenig Trainingsmaterial, das zum „overfitting“ führt.

Validierungs- und Trainingskurve verbessern sich bis am Ende des Trainings

Wenn beide Kurven bis zu den letzten Epochen (Trainingszyklen) leichte Verbesserungen feststellen lassen, ist das Netz noch nicht „austrainiert“. Optimal verbessert sich das Netz in den letzten 10-15 Epochen nur noch minimal bzw. gar nicht mehr. Wenn der Effekt der Verbesserung bis zum Ende anhält, sollte das Training nochmals mit mehr Epochen gestartet werden. Erfahrungsgemäss ist die Erkennung von austrainierten Netzen besser und zuverlässiger.

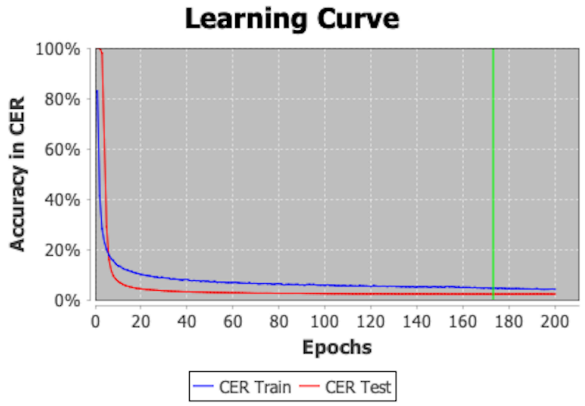

Abbildung 3: Lernkurve eines Netzes, das noch weiter austrainiert werden könnte.

Die Anwendung von Texterkennung, etwa mit Transkribus, ist ohne technische Vorkenntnisse problemlos erlernbar. Mit Rücksicht auf einige wenige Kniffe und mit basalen Kenntnissen der angewandten Algorithmen lassen sich grössere Dokumentenmengen sinnvoll und effizient aufbereiten.

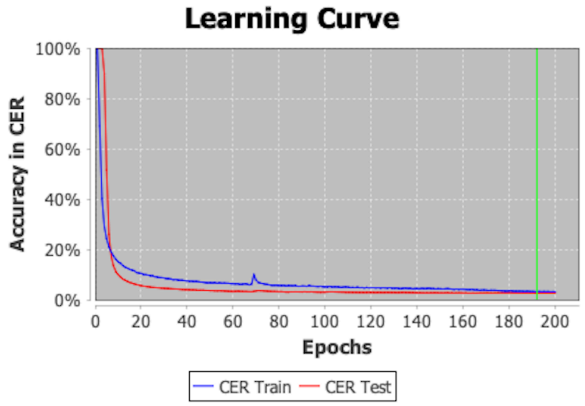

Abbildung 4: Lernkurve eines austrainierten Netzes mit genügend Epochen und ausreichend Trainingsmaterial.

Fußnoten

1. Im Rahmen eines Wettbewerbs an der ICFHR 2018 wurden nur noch Algorithmen basierend *machine learning* zur Erkennung der Handschriften eingesetzt, siehe: https://scriptnet.iit.demokritos.gr/competitions/10/viewresults/.
2. Eine Alternative zu Transkribus ist die Open Source Software Kraken, die ebenfalls auf der Basis neuronaler Netze Trainings individueller Handschriftenmodelle erlaubt. Für die Erkennung alter Drucke (gedruckt vor 1830) eignen sich auch Open Source Tools wie Tesseract. Siehe dazu den aktuellen Stand der Förderinitiative OCR-D, Neudecker u. a. 2019.
3. Aktuell ist der Zugang zur Trainingsfunktionalität nicht standardmässig gegeben. Per Mail (an email@transkribus.eu) wird die Möglichkeit aber rasch und unkompliziert gewährt.

Bibliographie

Ares Oliveira, Sofia / Seguin, Benoît / Kaplan, Frédéric (2018): dhSegment: A generic deep-learning approach for document segmentation. In: *Frontiers in Handwriting Recognition (ICFHR), 2018, 16th International Conference*. 7–12.

Hodel, Tobias (2018): Konsequenzen automatischer Texterkennung – Ein Aufriss zur Texterkennung mit Machine Learning. In: Vogeler, Georg (Hg.). *DHd 2018. Kritik der digitalen Vernunft Konferenzabstracts. Universität zu Köln 26. Februar bis 2. März 2018* Köln, 249–251. http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf.

Leifert, Gundram u. a. (2016): Cells in Multidimensional Recurrent Neural Networks. *Journal of Machine Learning. Res.* 17/97:1–97:37.

Muehlberger, Guenter u. a. (2019): Super Transforming scholarship in the archives through handwritten text recognition: Transkribus as a case study. *Journal of Documentation 75/5* , 954-976. https://doi.org/10.1108/JD-07-2018-0114

Neudecker, Clemens u. a. (2019): OCR-D: An end-to-end open source OCR framework for historical documents. *EuropeanaTech Insight* 13.

READ: Transkribus . https://read.transkribus.eu/transkribus/ [12.9.2019].

Quirós, Lorenzo (2017): *P2PaLA: page to PAGE layout analysis tookit.* https://github.com/lquirosd/P2PaLA.

Sahle, Patrick (2013): Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 2: Befunde, Theorie und Methodik. *Schriften des IDE 8* Bd. 2, Norderstedt: BoD. http://kups.ub.uni-koeln.de/5352/ [25.7.2014].

Sánchez, J.A. u. a. (2013): TranScriptorium: a European Project on Handwritten Text Recognition. In: Marinai, Simone / Marriott, Kim (Hg.). *ACM Symposium on Document Engineering DOCENG*. ACM, 227–228.

Schöch, Christoph (2017): Quantitative Analyse, in: Jannidis, Fotis / Kohle, Hubertus, Rehbein, Malte (Hg.). *Digital Humanities: Eine Einführung*. J.B. Metzler, Stuttgart, 279–298. https://doi.org/10.1007/978-3-476-05446-3_20