READing Handwritten Documents

Projekt READ und das Staatsarchiv Zürich auf dem Weg zur automatischen Erkennung von handschriftlichen Dokumenten

Tobias Hodel[1]

Recent advances in machine learning have led to substantial improvements in the recognition of layouts and handwriting in historical documents. The EU-funded READ project (https://read.transkribus.eu/) has developed cutting-edge Handwritten Text Recognition technology which allows scholars to automatically transcribe documents of any date, script or language.
This technology is freely available in the Transkribus platform, where users can train algorithms to recognize and search large collections written in single or multiple hands. The paper discusses the current state of the technology as well as use-cases for the recognition of handwriting. A key role is given to archives in the project in order to evaluate results and expectations with computer scientists.

Die historische Forschung braucht Quellen, um Bilder, Eindrücke und Einschätzungen vergangener Zeiten zu gewinnen. Diese Quellen gelten als Rohstoff der Wissenschaft und überdauerten in unterschiedlichsten Zusammenhängen die Zeit. Mittels Digitalisierung und durch die Aufarbeitung der Überlieferung in Archiven wurden in den vergangenen Jahrzehnten grosse Mengen von Dokumenten, insbesondere Texten, einfach zugänglich gemacht.

Aller Aufarbeitung zum Trotz ist der Zugriff auf die Dokumente nur bedingt gegeben. Eine vollständige Durchsuchbarkeit wird etwa erst möglich, nachdem die Texte bearbeitet worden sind. Für Drucke hat sich in den vergangenen Jahren die *optical character recognition* (OCR) durchgesetzt. Damit können mit hoher Genauigkeit Zeichen und Worte identifiziert werden. Noch immer ist die Technik jedoch fehleranfällig, was sofort auffällt, wenn Frakturtexte erkannt werden. Eine weit höhere Stufe der Komplexität

1 Wissenschaftlicher Mitarbeiter am Staatsarchiv des Kantons Zürich. Kontakt: tobias.hodel@ji.zh.ch. Alle URLs wurden letztmals am 27. 2. 2017 abgerufen.

wird erreicht, sobald nicht nur standardisierte Lettern erkannt werden sollen, sondern die Handschrift von Menschen.

Die Mehrheit der bis Mitte des letzten Jahrhunderts produzierten Schriftstücke liegt in dieser Form mit ihren individuellen Eigenheiten vor. Doch auch die menschliche Handschrift soll in Zukunft durch Algorithmen dahingehend erkennbar werden, dass eine weitere Verarbeitung und vor allem eine Durchsuchbarkeit der Texte ermöglicht wird. Die Chancen und Möglichkeiten, die sich aus einer solchen automatisierten Extraktion von Text aus Handschriften ergeben, sind für Wissenschaftler, die sich mit der Vormoderne beschäftigen, offensichtlich. Aber auch für die Neuzeit werden dadurch neue Quellenkorpora zugänglich und per Klick durchsuchbar.

Die Qualität der Volltexte wird vorwiegend von der Menge des zur Verfügung gestellten Trainingsmaterials abhängen, das in den nächsten Jahren für die Entwicklung der Technologien erzeugt wird. Das Staatsarchiv Zürich partizipiert in dem Zusammenhang an READ *(Recognition and Enrichment of Archival Documents),* einem wegweisenden Projekt, das im Rahmen von «Horizon 2020», dem Förderprogramm der Europäischen Union, bis Mitte 2019 die Technologie zur Einsatzreife in unterschiedlichen Kontexten entwickeln wird.

Im Rahmen der vorliegenden Vorstellung werden die technischen Grundlagen erörtert, die Projektpartnerschaften mit den Erinnerungsinstitutionen skizziert und künftig mögliche Weiterverarbeitungsschritte angetönt.

Technische Grundlagen der Handschriftenerkennung

Sowohl bei der *optical character recognition* (OCR) als auch bei der *handwritten text recognition* (HTR) basiert die Erkennung von Texten auf zwei Pfeilern: einerseits auf der Layout-Analyse, also der Identifikation, wo sich in einem zweidimensionalen Raum Schrift findet; andererseits auf der Erkennung dieser Zeichen als Buchstaben, Zahlen oder Satzzeichen und somit letztendlich als Text. Beide Schritte werden im Rahmen von READ von unterschiedlichen Forschungsteams mehrfach entwickelt. Somit entstehen im Lauf des Projekts nicht nur zwei, sondern mehr Sets von Algorithmen, die gegeneinander abgewogen und auch kombiniert werden können.

Im Fall der Layout-Erkennung wird versucht, Textregionen zu identifizieren und innerhalb dieser Regionen Linien und, wichtiger noch, *baselines* festzustellen. Unter *baseline* versteht man die imaginäre Linie, auf der eine Hand schreibt.[2] Basierend auf dieser Identifizierung, ist in einem zweiten Schritt die eigentliche Handschriftenerkennung möglich.

2 Zur automatisierten Layouterkennung siehe: STAMATOPOULOS Nikolaos, GATOS Basilis,

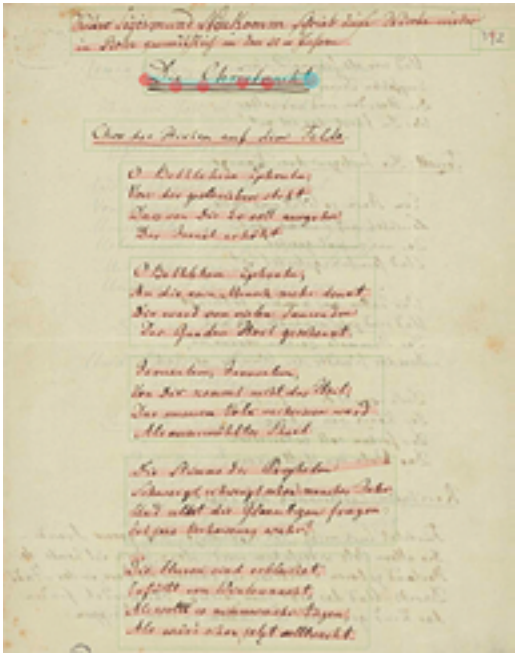

Abb. 1: Screenshot einer segmentierten Seite aus Transkribus. (NEUKOMM Sigismund, *Die Christnacht.* Universitätsbibliothek Basel, Autogr SarasinCh 392, https://www.e-manuscripta.ch/bau/doi/10.7891/e-manuscripta-13885)

HTR funktioniert, anders als OCR, nicht auf der Erkennung von einzelnen Zeichen, sondern auf der Analyse von Zeichenfolgen, womit sowohl ganze Zeilen als auch Worte gemeint sein können. Die erkannten Zeichenfolgen werden dann in unterschiedlicher Reihenfolge durch rekurrente neuronale Netze «gelesen» und Angaben zur Wahrscheinlichkeit der Lesung eines Zeichens werden gespeichert. Als Auswertung wird schliesslich entweder die wahrscheinlichste Lesung ausgegeben oder aber die Wahrscheinlichkeitstabelle (*confidence matrix*) für Suchanfragen zur Verfügung gestellt, um ein sogenanntes *keyword spotting* zu ermöglichen.[3]

Die Qualität der Erkennung mittels rekurrenter neuronaler Netze (RNN) basiert auf der Menge des Trainingsmaterials, das zur Verfügung steht, um einen Schrifttyp beziehungsweise die Eigenheiten eines Schreibers einzuüben.[4] Ab ungefähr 100 Seiten Trainingsmaterial wird eine Erkennung

«Goal-Oriented Performance Evaluation Methodology for Page Segmentation Techniques», in: Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR), 2015, 281–285 und Grüning, Tobias; Leifert, Gundram; Strauß, Tobias u. a.: A Two-Stage Method for Text Line Detection in Historical Documents, in: CoRR abs/1802.03345, 2018. Online: https://arxiv.org/abs/1802.03345.

3 Zu *keyword spotting* siehe TOSELLI Alejandro Hector et al., «HMM Word Graph Based Keyword Spotting in Handwritten Document Images», *Information Sciences,* 370–371 (2016), 497–518, DOI: 10.1016/j.ins.2016.07.063.

4 Zu den rekurrenten neuronalen Netzen siehe LEIFERT Gundram et al., «Cells in

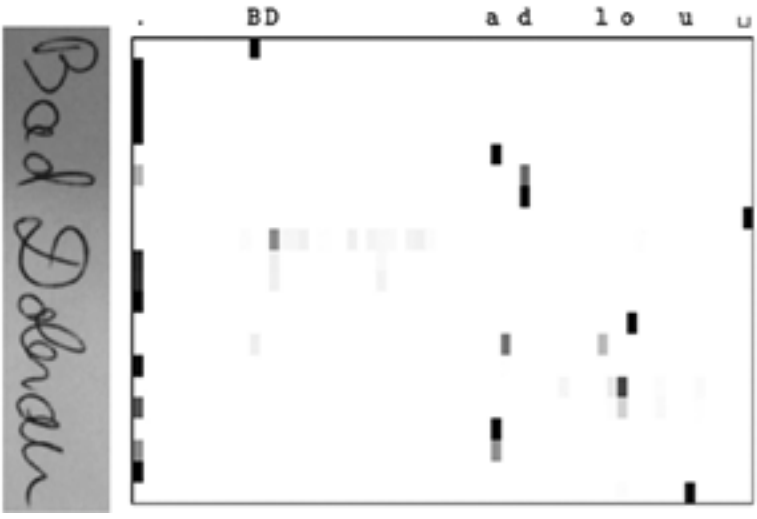

Abb. 2: GRÜNING Tobias, «Erkennungsmatrix – Confidence Matrix», Vortrag im Rahmen des Workshops *Digital Humanities und Editionswissenschaften,* Folie S. 14, http://www.hist.uzh.ch/dam/jcr:4e582cb5-940a-422f-bcb1-022110dd35fd/Gruning.pdf

im Bereich einer *character error rate* (CER) von 10 % möglich. Die CER wird durch den Vergleich der Erkennung mit der Vorlage von aufbereiteten (Test-)Seiten ermittelt, die nicht für das Training genutzt worden sind. Damit lässt sich die Zahl der Fehler pro 100 Zeichen eruieren.

Im Idealfall wird für einen Schreiber bei 1400 Seiten aufbereiteten Materials eine Erkennung im Bereich einer CER von 3 % möglich. Inwiefern eine gute (= CER < 10 %) Erkennrate – genügend Trainingsmaterial vorausgesetzt – für unterschiedliche aber ähnliche Hände, beispielsweise die Deutsche Kurrent des 19. Jahrhunderts, erreichen lässt, ist Teil der Forschungen in READ und wird im Rahmen des ständigen Austauschs zwischen den Projektpartnern über die Grenzen der Disziplinen hinaus diskutiert.[5]

Weiter verbessert wird die Erkennung über den Einbezug von Wörterbüchern, die übernommen oder eigens angelegt werden können. Grundsätzlich wird jedoch der Erkennung durch die neuronalen Netze

Multidimensional Recurrent Neural Networks», in: Journal of Machine Learning Research 17, 97:1-97:37 (2016), https://arxiv.org/abs/1412.2620v2.

5 MÜHLBERGER Günter, KAHLE Philip, COLUTTO Sebastian, «Handwritten Text Recognition (HTR) of Historical Documents as a Shared Task for Archivists, Computer Scientists and Humanities Scholars: The Model of a Transcription & Recognition Platform (TRP)», in: *HistoInformatics,* 2014,

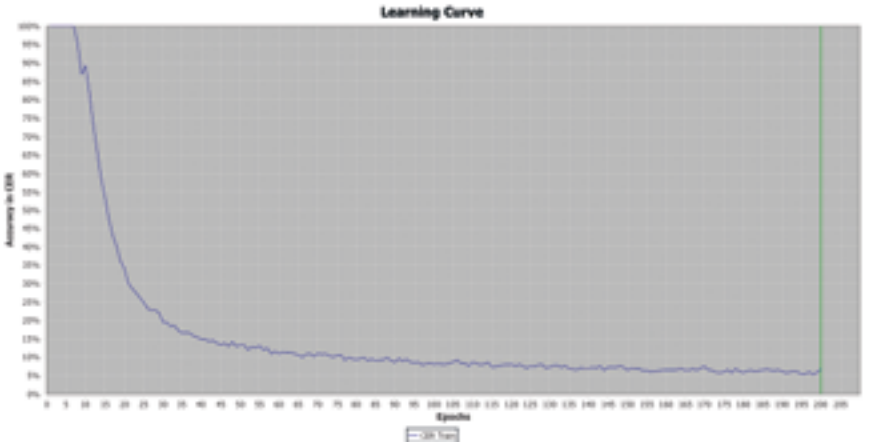

Abb. 3: Screenshot aus Transkribus. Entwicklung der Fehlerrate *(character error rate)* beim Training rekurrenter neuronaler Netze (200 Durchgänge) mit circa 400 Seiten Trainingsmaterial.

der Vorrang eingeräumt, damit Schreibvarianten nicht unnötigerweise korrigiert werden und erhalten bleiben; der sogenannten Hyperkorrektur wie sie häufig in OCR-Prozessen gefunden wird, kann somit ein Riegel geschoben werden.

Rolle der Erinnerungsinstitutionen

Im Gegensatz zu rein computerwissenschaftlich orientierten Forschungsprojekten bezieht READ Stakeholder aus den Geisteswissenschaften in den Entwicklungsprozess mit ein. Neben Geisteswissenschaftlern werden dabei vorwiegend Erinnerungsinstitutionen wie Bibliotheken und Archive angesprochen, die ihre Bestände nicht nur ins Netz stellen wollen, sondern den Nutzenden einen Zugriff auf Texte in Handschriften anbieten.
Im Rahmen der Kooperation mit READ wurde das Staatsarchiv Zürich beauftragt, eigenes Trainingsmaterial zur Verbesserung von Kurrenthandschriften herzustellen. Basierend auf Transkriptionen, die im Rahmen des Projekts TKR (Transkription und Digitalisierung der Kantonsratsprotokolle und Regierungsratsbeschlüsse des Kantons Zürich seit 1803)[6] erstellt wurden, erfolgten mehrere Trainings- und Evaluationsphasen. Aus den Erfahrungen

6 Für die Projektbeschreibung siehe http://www.staatsarchiv.zh.ch/internet/justiz_inneres/sta/de/ueber_uns/organisation/editio nsprojekte/tkr.html.

wird es möglich abzuschätzen, welche Textgüte mit spezifischem Training erreicht werden kann und welche Kosten im Prozess entstehen.

Gleichzeitig werden Weiterverarbeitungsschritte wie *document understanding* und *entity recognition* getestet.[7] Mit beiden Verfahren werden nach Möglichkeit Informationen identifiziert und aufbereitet, um Strukturen wie Titel oder Seitenzahlen sowie Orts- bzw. Personenangaben aus den erkannten Texten zu extrahieren.

In Zukunft werden Historikerinnen und Historiker dank der momentan entwickelten Technologien eine weitaus grössere Menge an Texten und Dokumenten berücksichtigen können, die in einer verbesserten Aufbereitung vorliegen, sodass eine trennschärfere Auswahl aufgrund einer breiteren Datengrundlage getroffen werden kann.

Die Software «Transkribus» ist frei verfügbar unter http://transkribus.eu. Weiterführende Informationen, ständige Aktualisierungen und Berichte zu den einzelnen Forschungsbereichen finden sich auf der Projekthomepage unter http://read.transkribus.eu

Das Projekt wird im Rahmen des Forschungs- und Innovationsprogramms «Horizon 2020» der Europäischen Union (Grant Agreement No 674943) gefördert.

Tobias Hodel
is postdoc at the state archives of Zurich and responsible for project READ, working on handwritten text recognition. He pursued a PhD in history (defended in 2016) about archival practices in Königsfelden Abbey. Hodel is in charge of a digital edition project (funded by the Swiss National Science Foundation) and the e-learning platform Ad fontes.

7 Zum *Document Understanding* siehe DEJAN Hervé, «Document Analysis and Layout Using Sequential Pattern Mining Techniques», in: *XRCE Blog,* 23. 1. 2017, https://web.archive.org/web/20170127104937/www.xrce.xerox.com/Blog/Document-Analysis-and-Layout-Using-Sequential-Pattern-Mining-Techniques.