Handschriftenerkennung für historische Schriften. Die Transkribus Plattform

Günter Mühlberger und Tamara Terbul

Historische Handschriften des Mittelalters, der frühen Neuzeit oder des 19. und 20. Jahrhunderts automatisiert erkennen und durchsuchen – das wäre noch vor wenigen Jahren als unrealistisch eingestuft worden. Die technologischen Fortschritte beim maschinellen Lernen, sowie der Bild- und Sprachverarbeitung machen jedoch genau das möglich. Mit der Forschungsplattform Transkribus, die im Rahmen des EU Projekts READ entwickelt wurde, steht nun diese Technologie jedem Benutzer frei zur Verfügung. Aber nicht nur das: Für Archive und Bibliotheken eröffnet das sogenannten Keyword Spotting Verfahren die Möglichkeit handschriftliche Dokumente nun auch mit wesentlich größerer Genauigkeit durchsuchbar machen zu können als dies bisher mit den Methoden der Volltextsuche möglich war.

Handwriting recognition for historical manuscripts
It would have been considered unrealistic just a few years ago to imagine one could automatically recognize and search through histrical manuscripts from the Middle Ages, from the Early Modern Period or from the 19th and 20th century would have been considered unrealistic. Yet, technological advances in machine reading and in image and language processing have rendered it possible. With Transkribus, the research platform developed in the context of the EU-project READ, this technology is now generally available for free. And what is more: Keyword Spotting allows archives and libraries to search hand-written documents with a much higher degree of precision than with present methods of full text search.

Transkribus als Plattform

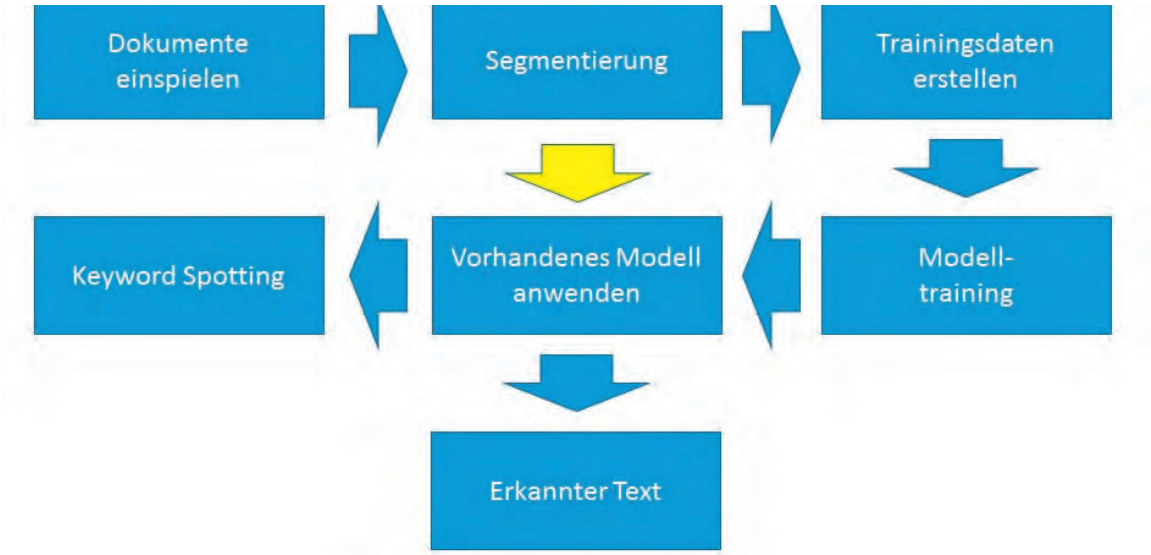
Maschinenlernende Verfahren benötigen Daten – und je mehr Daten zur Verfügung stehen, desto besser die Ergebnisse. Im Fall historischer Schriften sind das korrekte Transkriptionen und die dazugehörigen Bildausschnitte einer einzelnen Zeile. Diese Trainingsdaten zentral zu sammeln und neue Modelle auf den aggregierten Daten zu generieren, um diese Modelle dann wieder anderen Benutzern zur Verfügung stellen zu können, das war der Grundgedanke bei der Konzeption der Transkribus Plattform. Denn obwohl alle Dokumente in Transkribus grundsätzlich privat sind und nur für den jeweiligen Benutzer, der die Daten ins System eingebracht hat, sichtbar sind, können doch die daraus generierten Modelle bedenkenlos geteilt werden. Mehr als 10.000 Benutzer aus der ganzen Welt haben sich bereits in der Plattform registriert – und mehr als 400 verschiedene Handschriftenmodelle wurden schon in den unterschiedlichsten Sprachen und Alphabeten trainiert. Im dritten Jahr der Plattform werden die in der Plattform erzielten Netzwerkeffekte nun erstmals sichtbar. So stehen inzwischen für mittelalterliche lateinische Schriften, aber auch historisches Englisch, Holländisch, Deutsch oder Finnisch viele unterschiedliche Daten zur Verfügung, die es oftmals erlauben, bereits auf ein vorhandenes Modell zurückzugreifen. Damit wird deutlich, dass die Digitalisierung im Archiv- und Bibliotheksbereich nur dann ihr gesamtes Potential entfalten kann, wenn möglichst viele der an diesem Prozess beteiligten Archive, Bibliotheken, Geisteswissenschaftler, Studenten, Familienforscher und Computerwissenschaftler die Gelegenheit erhalten, über die engeren Fachgrenzen hinweg zusammenarbeiten zu können. Bis zum Ende des Projekts im Juni 2019 soll daher Transkribus als eigenständige Forschungsinfrastruktur etabliert werden. Derzeit wird dabei das Modell einer European Cooperative Society („Genossenschaft“) favorisiert, das eine eigenständige Unternehmensform darstellt und es den Mitgliedern ermöglicht Geschäfte mit der Kooperative abzuschließen und doch gleichzeitig auch Miteigentümer zu sein. Der besondere Charme einer Kooperative liegt dabei in ihrer demokratischen Ausrichtung aber auch in einem klaren Bekenntnis zu einem gewinnorientierten Modell – wobei die erzielten Gewinne nicht einer einzelnen Institution zugutekommen, sondern ein „Gewinn für alle“ erbracht wird.[1]

Der Weg zum Handschriftenmodell

Die Erstellung eines Schriftenmodells für historische Druck- oder Handschriften ist einfach[2] und besteht aus den folgenden Schritten: Upload der digitalisierten Dokumente in die Plattform, Segmentierung, Transkription und Training. Sollte bereits ein Modell vorhanden sein, dann kann dieses direkt angewendet werden. Als Output gibt es einerseits den automatisiert erstellten Text und andererseits die Möglichkeit zum Keyword Spotting.
Die Plattform verfügt derzeit über einen dedizierten Speicherplatz von 100 TB, sodass in den nächsten

1 Vgl. den Leitspruch der „Genossenschaften in Deutschland“. Online unter: http://genossenschaften.de/

2 Genaue Anleitungen finden sich in den Benutzerinstruktionen im Transkribus Wiki: https://transkribus.eu/wiki/index.php/How_to_Guides



Transkribus: Workflow

Jahren rund 20 Mill. Dateien in Transkribus gespeichert werden können.

Nach erfolgreichem Upload der Dokumente besteht der erste Schritt in der (automatisierten) Segmentierung einer Seite in Textblöcke und Textzeilen. War dieser Schritt bis Ende 2017 noch mit viel manueller Arbeit verbunden, so konnten hier die größten Fortschritte im READ Projekt erzielt werden. Die Layoutanalyse der Universität Rostock (CITlab Team) findet nunmehr mit großer Genauigkeit auch in komplexen Dokumenten alle Textzeilen.[3]

Im Anschluss an die Segmentierung der Dokumente können die ersten Trainingsdaten erstellt werden. Dafür gibt es zwei Möglichkeiten: einerseits durch eine manuelle Transkription des Textes oder mittels eines automatisierten Verfahrens, das bereits vorhandene Transkriptionen mit dem Bild verknüpft. Bei der manuellen Transkription ist grundsätzlich eine buchstabengetreue Transkription von Vorteil, typische Abkürzungen oder Verschleifungen können jedoch durchaus normalisiert eingegeben werden. Sollten paläografisch hochwertige Transkriptionen benötigt werden, so können alle Uni-Code Zeichen mittels eines virtuellen Keyboards eingegeben werden.

Bei der automatisierten Erstellung von Trainingsdaten wird dem Umstand Rechnung getragen, dass in vielen Archiven, Bibliotheken und Forschungsgruppen bereits hochwertige Transkriptionen vorhanden sind. Es wurde daher ein Tool entwickelt, das vorhandene Transkriptionen mit dem entsprechenden Bild einer Seite automatisiert zusammenführt. Auf diese Weise können tausende von Seiten, die z.B. im Rahmen einer (digitalen) Edition entstanden sind, mit geringem Aufwand für das Training umfassender Modelle genutzt werden.[4]

Zusätzlich zum Experteninterface kann die Transkription auch mit einem Webinterface vorgenommen werden. Dieses richtet sich besonders auch an ehrenamtliche Mitarbeiter, die im Rahmen eines „Crowd-Sourcing“ Projekts zur Erstellung von historischen Texten beitragen möchten.

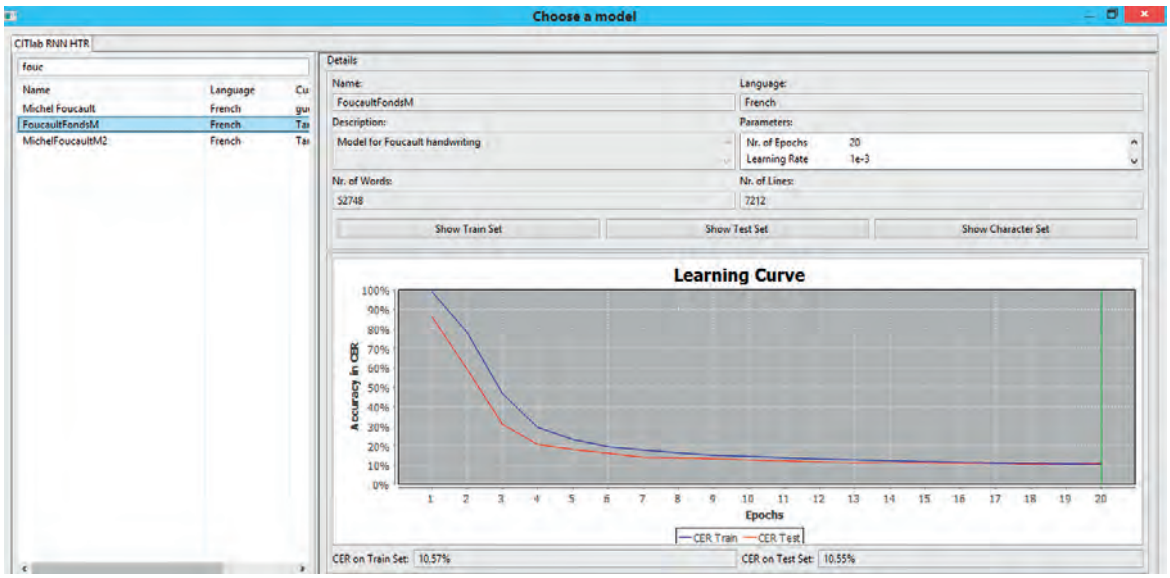
Das Training des Modells

Wurden die einzelnen Schritte durchgeführt, kann nun das erste Modell trainiert werden. Es spielt dabei keine Rolle, ob es sich um eine Handschrift oder eine Druckschrift handelt. Für die Durchführung des Trainings ist eine bestimmte Anzahl an transkribierten Wörtern nötig. Wie immer beim maschinellen Lernen gilt: je mehr Daten, desto besser. Grundsätzlich kann man jedoch davon ausgehen, dass mit 5000 korrekt transkribierten Wörtern bereits ein erstes Training Sinn macht. Für wirklich leistungsfähige Modelle werden jedoch üblicherweise wesentlich mehr Wörter benötigt, etwa 20.000 bis 30.000 – was ungefähr 100 bis 150 Seiten entspricht.

Beim Training selbst sind nur wenige Parameter zu beachten, in den allermeisten Fällen führen die Standardwerte zu guten Ergebnissen. Nach dem Training werden die Lernkurve und die Character Error Rate (CER) den Nutzern angezeigt. Die CER wird in Bezug auf die zwei, vom Benutzer ausgewählten Sets, dargestellt: dem Trainingsset und dem Testset. Die CER in Bezug auf das Trainingsset neigt gerne zur „Überanpassung“, d.h. das Neuronale Netz, das hier zum Einsatz kommt, lernt die Daten „auswendig“ und erzielt

3 Tobias Grüning, et al.: A Robust and Binarization-Free Approach for Text Line Detection in Historical Documents. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)2017 DOI: 10.1109/ICDAR.2017.47

4 Gundram Leifert, Tobias Strauß, Roger Labahn: D7.20 Model for Semi- and Unsupervised HTR Training P2. How to get a good HTR without expensive ground truth production. 2017. https://read.transkribus.eu/wp-content/uploads/2017/12/Del_7.20.pdf



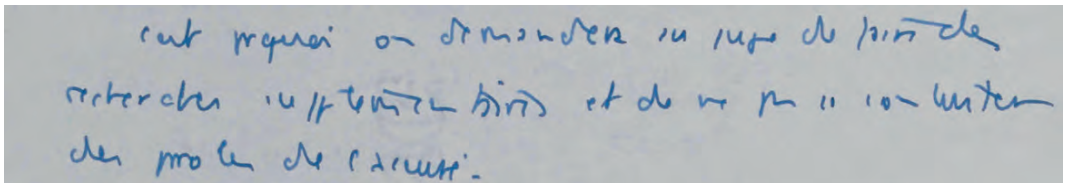
Lernkurve, Parameter und Fehlerquote eines HTR Modells für Michele Foucault

dadurch besonders gute Ergebnisse. Wendet man es hingegen auf neue Daten an, dann sind die Ergebnisse oftmals deutlich schlechter. Beim Testset besteht diese Gefahr nicht, da diese Daten nicht in das Training eingehen. Hier ist jedoch besonders auf die Auswahl zu achten: Enthält das Testset Dokumente, die nicht oder nur wenig repräsentativ für das Trainingsset bzw. für die später zu erkennenden Seiten sind, dann ist die Aussagekraft klarerweise gering. Als Faustregel kann gelten, dass die Ergebnisse sowohl repräsentativ als auch robust sind, wenn die Werte, die gegen das Trainings- und das Testset gemessen werden, nicht zu weit voneinander abweichen.

Die Texterkennung lieferte beispielsweise bei dem unten dargestellten Modell für die Handschrift von Michele Foucault mit 52.748 Wörtern eine CER am Trainingsset von 10,57% und 10,55% am Testset.

Aber selbst diese Werte repräsentieren nicht mehr den „State-of-the-art". Neuere Implementierungen neuronaler Netze reduzieren diese Fehlerquote um mind. 50% – sodass man davon ausgehen kann, dass die obigen Trainingsdaten ausreichen, um Werte unter 5% CER bei dieser Handschrift zu erreichen.[6] Das bedeutet, dass nur noch jedes 20. Zeichen falsch erkannt wird, bei einer Handschrift, die auch unter Experten als durchaus „herausfordernd" eingestuft wird. Für Druckschriften des 16. – 18. Jahrhunderts sind hingegen Werte unter 1-2% CER als durchaus realistisch anzusehen.

Ist erst einmal ein Modell trainiert, so wird dieses direkt für den Benutzer verfügbar. Der Benutzer wählt das Modell und ein Dokument aus und führt eine Zeilenfindung sowie die Erkennung mit dem erstellten Modell durch. Die Anwendung auf neue Seiten benötigt derzeit weniger als eine Minute pro Seite, auch dieser Vorgang wird durch die Einführung neuer Technologie noch wesentlich beschleunigt werden.



Handschrift Michel Foucaults[5] (fonds Michel Foucault, BnF, NAF 28730 boîte 1, folio n°157)

Durch Zugabe von weiteren rund 35.000 Wörtern konnte die CER beim Trainingsset auf 8,40% und im Testset auf 9,31% verbessert werden

C est prquoi on demandera au juge de faire des rechercbes supplementaires et de ne pas se conteiter des paroles de l accure.	C est prquoi on demandera au juge de faire des recherches supplémentaires et de ne pas se contenter des paroles de l accusé.

Vergleich erkannter mit korrektem Text bei ca. 10% CER

5 Dieser unpublizierte Textausschnitt wurde uns freundlicherweise von Vincente Ventresque, Projektleiter des Forschungsprojekts ANR Foucault Fiches de lecture zur Verfügung gestellt.

6 Vgl. etwa: J. A. Sánchez, V. Romero, A. H. Toselli, M. Villegas, E. Vidal, „ICDAR2017 Competition on Handwritten Text Recognition on the READ Dataset" 2017, 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 1383-1388. Online: doi: 10.1109/ICDAR.2017.226

Keyword Spotting

Sobald in Transkribus ein Dokument mit einem HTR Modell erkannt wurde, kann es auch mit dem sogenannten Keyword Spotting Verfahren durchsucht werden. Keyword Spotting bedeutet im Gegensatz zu einer herkömmlichen Volltextsuche, dass nicht innerhalb des Volltextes gesucht wird, sondern eine Stufe zuvor angesetzt wird: Gesucht wird in einer Konfidenzmatrix, die für jeden Punkt auf einer Seite einen Wahrscheinlichkeitswert angibt, mit der darin ein Buchstabe aus dem zugrundeliegenden Alphabet repräsentiert wird.[7] Statt sich also „festzulegen" gibt das Netz nur an, dass es mit einer bestimmten Sicherheit davon ausgeht, dass sich an einer bestimmten Stelle im Bild der Buchstabe a, b, c, d, etc. verbirgt. Dies hat zur Folge, dass selbst wenn ein Wort nicht gut leserlich ist oder das trainierte Modell keine guten Ergebnisse erzielt, es vom Keyword Spotting trotzdem gefunden werden kann. Das folgende Beispiel zeigt die Leistungsfähigkeit der KWS:

Beispiel für Keyword Spotting

Die automatisierte Transkription für diese Schrift ergab: **Josrich Mitrltser**. Während jede Volltextsuche an dieser Aufgabe scheitern würde, findet die KWS den gesuchten Begriff – hier der Familienname „Mitterlehner" – mit einer ausreichenden Sicherheit von 0,16 (0=unsicher, 1=sicher). Aber nicht nur das: Mit dem Verfahren werden auch noch Schwierigkeiten, wie sie für historische Schreibweisen und Varianten typisch sind, gelöst. So liefert etwa die Suche nach „Josef Mitterlehner" (statt: Joseph) ebenfalls das gewünschte Ergebnis, einzig die Sicherheit verringert sich auf 0,11.

Die Vorteile für Archive, Bibliotheken, Historiker, Familienforscher und Genealogen liegen auf der Hand. Auch handschriftliche Dokumente können nun erstmals nach allen beliebigen Wörtern durchsucht werden – auch wenn die automatisierte Transkription im konkreten Fall einen wenig hilfreichen Text produziert. Im Rahmen des READ Projekts werden 2018 mindestens zwei größere Demonstrationsprojekte durchgeführt, bei denen das Keyword Spotting Verfahren für Archive und Bibliotheken im Mittelpunkt stehen wird. Einmal soll damit der gesamte Bestand von Jeremy Bentham zugänglich gemacht werden. Es handelt sich um knapp 100.000 Seiten, die auf diese Weise erschlossen werden sollen. Zum anderen sollen damit knapp 1 Mill. Seiten von Gerichtsurteilen aus dem 19. Jahrhundert des Finnischen Nationalarchivs für die Öffentlichkeit zugänglich gemacht werden. In beiden Fällen ist davon auszugehen, dass bereits einige hundert Seiten korrekt transkribierte Texte für das Training umfassender Modelle ausreichend sein werden.

Zusammenfassung

Die Möglichkeit, Spezialmodelle für bestimme Hand- und Druckschriften trainieren und anwenden zu können, ist besonders für Geisteswissenschaftler von Vorteil. Auf diese Weise kann oftmals rascher eine Transkription angefertigt werden, als dies durch das reine Abtippen möglich wäre – das obige Beispiel der Foucault-Edition zeigt dies eindrücklich. Hinzu kommt, dass eine gute Erkennung auch sinnverstehendes Lesen ermöglicht, ohne dass z.B. die zugrundeliegende Schrift in allen Einzelheiten für den Benutzer lesbar sein muss. Für Archive und Bibliotheken liegt der wesentliche Vorteil der neuen Technologie jedoch ohne Zweifel im Keyword Spotting Verfahren begründet. Sie können damit die Bedürfnisse sowohl von Wissenschaftlern als auch Familienforschern oder der breiten Öffentlichkeit nach einer einfachen und umfassenden Volltextsuche innerhalb einer Sammlung rasch und einfach befriedigen. Die Transkribus Plattform stellt für beide Szenarien die geeigneten Tools zur Verfügung und ermöglicht eine gemeinsame Nutzung der Daten über die Grenzen der jeweiligen Institutionen und Projekte hinweg.

Referenzen:
Transkribus Plattform: http://transkribus.eu/
READ Projekt: http://read.transkribus.eu/

Mag. Dr. Günter Mühlberger
Projektkioordinator READ
Universität Innsbruck
Digitalisierung und elektronische Archivierung (DEA)
Innrain 52
A-6020 Innsbruck
guenter.muehlberger@uibk.ac.at

Tamara Terbul, MA
Leopold-Franzens-Universität Innsbruck
Forschungszentrum „Digital Humanities"
tamara.terbul@uibk.ac.at

7 Hier wird nur auf das KWS – Query by String eingegangen. Zu anderen Verfahren vgl. etwa: Joan Puigcerver; Alejandro H. Toselli; Enrique Vidal, „ICDAR2015 Competition on Keyword Spotting for Handwritten Documents" 2015, 13th International Conference on Document Analysis and Recognition (ICDAR) 2015 doi: 10.1109/ICDAR.2015.7333946