Automatic Text and Feature Recognition: Mit READ Werkzeugen Texte erkennen und Dokumente analysieren

Hodel, Tobias
tobias.hodel@uzh.ch
Staatsarchiv des Kantons Zürich, Schweiz

Diem, Markus
diem@cvl.tuwien.ac.at
Computer Vision Lab, TU Wien

Oliveira Ares, Sofia
sofia.oliveiraares@epfl.ch
Digital Humanities Lab, EPF Lausanne

Weidemann, Max
max.weidemann@uni-rostock.de
Citlab, Universität Rostock

Dank *machine learning* und *computer vision* ist seit wenigen Jahren die automatisierte Handschriftenerkennung möglich. Obwohl aktuell einzelne Handschriften bzw. sehr ähnliche Handschriftentypen noch trainiert werden müssen, wird es in absehbarer Zeit allgemeine Modelle geben, die Rohtranskriptionen mit einer Fehlerquote unter 10% ausgeben. Paläographische Kenntnisse werden vor allem zur Korrektur und kritischen Begutachtung der Technik nötig sein.

Im Rahmen des Projekts READ (Recognition and Enrichment of Archival Documents) werden unterschiedliche Aufgaben der Automatisierung (weiter-)entwickelt, um qualitativ gute Ergebnisse mit optimalem Ressourceneinsatz zu erhalten. Ein speziell dafür entwickeltes Tool ist die Software Transkribus und die Transkribus Weboberfläche (für Transkription, Tagging/Annotation und Korrektur in der Layouterkennung). Beide Ansätze verkoppeln auf unterschiedliche Weise die Arbeit von Expertinnen und maschinelle Erkennleistung. Software und Webservice sind frei verfügbar unter www.transkribus.eu . Darüber hinaus wurden im Rahmen von READ weitere Extraktions- und Annotationsmöglichkeiten entwickelt, die im Workshop zusammen mit Transkribus vorgestellt und durch die Teilnehmenden mit eigenen oder zur Verfügung gestellten Dokumenten getestet werden können.[1]

Transkribus unterstützt alle Prozesse vom Import der Bilder über die Identifikation der Textblöcke und Zeilen, die zu einer detaillierten Verlinkung zwischen Text und Bild führt, sowie die Transkription und Annotation der Handschrift bis zum Export der gewonnen Daten in standardisierten Formaten. Darüber hinaus wurden aber noch weitere Tools und Algorithmen entwickelt, die zur Erkennung von graphischen Features genutzt werden können und Tabellen als solche aufbereiten.

Transkribus als Arbeitsumgebung

Die Erkennung von Texten bedingt den Upload digitaler Bilder und Prozessierung mit Layouterkennungswerkzeugen. Upload und Layoutanalyse können grosse Batches verarbeiten. Die Nachkorrektur von Layoutanalyse ist nur noch in wenigen Fällen nötig.

Je nach Einsatzzweck könne Dokumente entweder automatisch mit bereits bestehenden ATR-Modellen (Automatic Text Recognition) erkannt oder händisch Transkriptionen erstellt werden. Um im erkannten Text und Variantenlesungen (sog. Keywordspotting) zu suchen reicht in den meisten Fällen die Anwendung von bestehenden Modellen.

Einzig im Umgang mit Tabellen sind weiterhin diverse manuelle Schritte möglich, um eine hochwertige Identifikation zu gewährleisten. Der Workshop wird einen Fokus auf die Bearbeitung und Erkennung von Tabellenstrukturen legen, die halbautomatisch erfolgen kann.

Korrekter Text – entweder durch Transkription oder Korrektur von erkanntem Text entstanden – kann danach zum Training von Handschriftenmodellen verwendet werden. Im Rahmen des Workshops wird das Trainieren von Handschriftenmodellen demonstriert und kann durch die Teilnehmenden selbst ausgetestet werden.

Aufbauend auf den Transkriptionen ist es möglich Entitäten (Personen, Orte, Verweise) auszuzeichnen und textuelle Annotationen (Titel, Marginalien, Fussnoten) innerhalb des Textes, aber auch darüber hinaus für Einzeldokumente und ganze Dokumentenbestände anzulegen. Visuelle Features wie Seitenzahlen, Titel oder Marginalien lassen sich nach der Auszeichnung als Strukturmodelle trainieren und können für die Erkennung von grösseren Dokumentenmassen verwendet werden. Die Vorgehensweise wird im Rahmen des Workshops vorgeführt und kann selbst nachvollzogen werden. Daneben ist auch die Anreicherung der Dokumente mit *named entities* (Personen, Orten und Organisationen) möglich, sodass simple digitale Editionen grösstenteils in Transkribus erstellt werden können.

Ausgabeformate

Für den Export stehen unterschiedliche Formate und Ausgabeformen zur Verfügung. So ist es möglich XML-Dateien zu exportieren, die den Vorgaben der TEI entsprechen (auch ist es möglich die Standardumformung abzuändern und den eigenen Bedürfnissen anzupassen). Weiter sind auch Ausgaben als Druckdaten (PDF) oder zur Weiterbearbeitung für Textverarbeitungsprogramme (DOCX, TXT) implementiert. Schließlich ist auch ein Export im PAGE-Format (zur Anzeige in Viewern für OCR gelesene Dokumente, Pletschacher, 2010) sowie als METS (Metadata Encoding and Transmission) möglich.

Zielpublikum

Die Plattform Transkribus ist für unterschiedliche Gruppen konzipiert. Einerseits für Geisteswissenschaftler*innen, die selbst Transkriptionen und Editionen historischer Dokumente erstellen möchten. Andererseits richtet sich die Plattform an Archive, Bibliotheken und andere Erinnerungsinstitutionen, die handschriftliche Dokumente in ihren Sammlungen aufbewahren und ein Interesse an der Suchbarmachung des Materials haben. Angesprochen werden sollen auch Studierende der Geistes-, Archiv- und Bibliothekswissenschaften mit einem Interesse an der Transkription historischer Handschriften.

Das Ziel, eine robuste und technisch hochstehende Automatisierung von Layout- und Handschriftenerkennung, lässt sich nur durch die enge Zusammenarbeit zwischen Geisteswissenschaftler*innen und Informatiker*innen sowie anderen Computerspezialist*innen mit unterschiedlichen Voraussetzungen und Ansprüchen an Datenqualität und Herstellung von Transkriptionen erreichen. Die Algorithmen werden somit nicht nur bis zu einem Status als *proof-of-concept* erarbeitet, sondern bis zur Praxistauglichkeit verfeinert und in größeren Forschungs- und Aufbewahrungsumgebungen getestet und verbessert. Die Informatiker*innen sowie Personen aus angrenzenden Fächern sind entsprechend ebenfalls ein wichtiges Zielpublikum, wobei bei ihnen weniger die Nutzung der Plattform als das Beisteuern von Software(teilen) anvisiert wird.

Die Speicherung der Dokumente erfolgt in der Cloud, gehostet auf Servern der Universität Innsbruck. Die importierten Daten bleiben auch während der Bearbeitung unverändert im Dateisystem liegen und werden ergänzt durch METS und PAGE XML. Alle bearbeiteten Dokumente und Daten bleiben somit in den unterschiedlichen Bearbeitungsstadien nicht nur lokal verfügbar, sondern können für andere Transkribusnutzerinnen und -nutzer freigegeben werden. Dank elaboriertem *user-management* ist die Zuteilung von Rollen möglich.

Die eingespeisten Dokumente und Daten bleiben privat und vor dem Zugriff Dritter geschützt. Von Projektseite können vorgenommene Arbeitsschritte zwecks besserem Verständnis der ausgeführten Arbeiten und letztlich der Verbesserung der Produkte ausgewertet werden.

Die Erkennprozesse werden serverseitig durchgeführt, sodass die Ressourcen auf den lokalen Rechnern nicht strapaziert werden. Transkribus ist mit JAVA und SWT programmiert und kann daher plattformunabhängig (Windows, Mac, Linux) genutzt werden.

Ein- und Ausblicke im Workshop

Der Workshop richtet sich sowohl an Geisteswissenschaftler*innen als auch an Computerwissenschaftler*innen, wobei vorwiegend die Tools und Möglichkeiten von Transkribus präsentiert werden.

Drei zentrale Forschungsaspekte aus READ können im Rahmen des Workshops neben Transkribus *hands-on* ausgetestet werden:

1. Max Weidemann: Das Training von Handschriftenmodellen (HTR+);
2. Sofia Ares Oliveira (*in English*): Identifikation von visuellen Features mit dh-segment;
3. Markus Diem: Aufbereitung und Erkennung von Tabellen mit Transkribus und nomacs.

Programm/Ablauf des Workshops

- Begrüssung und Einführung in READ und Transkribus :45’
- Kurze Beschreibungen der vermittelten Forschungsaspekte (je 15’): 45’
- Kaffeepause: 30’
- Arbeit in Kleingruppen am jeweiligen Forschungsaspekt: 60’ (nach ca. 40 Minuten besteht die Möglichkeit die Gruppe zu wechseln)
- Diskussion der Resultate, weiterer Ausblick und Evaluation: (15-)30'

Nach Interesse der Teilnehmenden können während der Gruppenarbeit weitere Tools und Ansätze, die im Rahmen von READ entwickelt wurden, kurz diskutiert werden: 1. Matching von Text und Bild (bspw. aus bestehenden Transkriptionen), 2. Transkribus Learn (e-Learningumgebung), 3. Crowdsourcing-Infrastruktur, 4. ScanTent und DocScan (Fotografieren von Dokumenten mit Android App).

Während des gesamten Workshops stehen vier wissenschaftliche Mitarbeitende des Projekts für Fragen und Auskünfte zur Verfügung.

Tobias Hodel nimmt bereits im Vorfeld gerne Dokumente oder Projektideen an, damit sich die Veranstalter bereits vor dem Workshop Gedanken zu möglichen technischen Umsetzungen machen können.

Das Projekt READ und somit die Weiterentwicklung von Transkribus werden finanziert durch einen Grant der Europäischen Union im Rahmen des Horizon 2020 Forschungs- und Innovationsprogramms (grant agreement No 674943).

Zahl der möglichen Teilnehmerinnen und Teilnehmer: Max. 30 Personen.

Benötigte technische Ausstattung: Beamer und Whiteboard.

Teilnehmende: Eigener Rechner (wenn möglich Installation von Transkribus; Hilfe zur Installation von Transkribus wird 15 Minuten vor der Veranstaltung angeboten).

Rückfragen bitte an tobias.hodel@ji.zh.ch

Kontaktdaten aller Beitragenden (inkl. Forschungsinteressen)

Sofia Ares Oliveira, École Polytechnique de Lausanne, CDH-DHLAB, INN 116 / Station 14 / CH-1015 Lausanne / Switzerland; sofia.oliveiraares@epfl.ch (Electrical engineering, signal processing, computer vision).

Markus Diem, Technische Universität Wien, Institute of Computer Aided Automation Computer Vision Lab, Favoritenstr. 9/183-2, A-1040 Vienna, Österreich; diem@caa.tuwien.ac.at (Computer Vision, Document Analysis, Layout Analysis/Page Segmentation, Cluster Analysis, Automated Flow Cytometry Analysis).

Tobias Hodel, Staatsarchiv des Kantons Zürich, Winterthurerstrasse 170, CH-8057 Zürich, Schweiz; tobias.hodel@ji.zh.ch (Digital Humanities; Automatic Text Recognition; eArchiving; Information Retrieval).

Max Weidemann, Institut für Mathematik, Ulmenstraße 69, Universität Rostock, 18051 Rostock, Deutschland; max.weidemann@uni-rostock.de; (Deep Learning, Information Retrieval und Natural Language Processing).

Fußnoten

1. Einführend siehe die online Tutorials: https://read.transkribus.eu/transkribus/ . Als *hands-on* Anleitung wird der Beitrag von Martin Prell empfohlen: »ps: ich bitt noch mahl umb ver gebung meines confusen und üblen schreibens wegen« – Frühneuzeitliche Briefe als Herausforderung automatisierter Handschriftenerkennung. Online: https://doi.org/10.22032/dbt.34849.

Bibliographie

Leifert, Gundram / Strauß, Tobias / Grüning, Tobias / Wustlich, Welf / Labahn, Roger (2016): *„Cells in Multidimensional Recurrent Neural Networks"* in: *Journal of Machine Learning Research* 17, 1-37.

Prell, Martin (2018): *„»ps: ich bitt noch mahl umb ver gebung meines confusen und üblen schreibens wegen« – Frühneuzeitliche Briefe als Herausforderung automatisierter Handschriftenerkennung"*. Online: https://doi.org/10.22032/dbt.34849.

READ (2018): *„Tutorials and How To Guides"*. Online: https://read.transkribus.eu/transkribus/.