Automatic Text Recognition: Mit Transkribus Texterkennung trainieren und anwenden

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

Strauß, Tobias
tobias.strauss@uni-rostock.de
Universität Rostock CITLab

Diem, Markus
diem@caa.tuwien.ac.at
Vienna University of Technology, Computer Vision Lab

Die Aufbereitung und Erkennung von handschriftlichen Dokumenten oder von speziellen Druckschriften ist sowohl für Menschen als auch für Computeralgorithmen eine (technische) Herausforderung. Die Bearbeitung von schriftlichem, insbesondere handschriftlichem Material aber auch früher Drucke wird bislang von spezialisierten Experten durchgeführt, um technisch und qualitativ hochstehende Resultate aus historischen Dokumenten zu erhalten. Zur Erstellung hochwertiger Editionen sind hilfswissenschaftliche Kenntnisse (Paläographie, Editorik), historisches Kontextwissen und technisches Know-how gefragt.

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 (öffentliche Vorstellung im November 2017). Beide Ansätze verkoppeln auf unterschiedliche Weise die Arbeit von Expertinnen und maschinelle Erkennleistung. Software und Webservice sind frei verfügbar unter www.transkribus.eu. Im Workshop wird Transkribus vorgestellt und kann durch die Teilnehmenden mit eigenen oder zur Verfügung gestellten Dokumenten getestet werden.

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.

Workflow in Transkribus

Um Texte zu transkribieren oder zu edieren, müssen digitale Bilder hochgeladen und danach mit Layouterkennungswerkzeugen bearbeitet werden. Die Analyse des Layouts kann automatisiert geschehen, wobei die manuelle Kontrolle und falls nötig die Nachbearbeitung im Moment noch sinnvoll ist.

Dokumente können entweder automatisch mit bereits bestehenden ATR-Modellen (Automatic Text Recognition) erkannt werden oder die Transkription erfolgt händisch und kann danach zum Training neuer Modelle genutzt werden. Insbesondere für die Bearbeitung großer Dokumentenkorpora, die in ähnlichen Handschriften verfasst wurden, lassen sich bereits heute Effizienzgewinne und Vereinfachungen erzielen.

Aufbauend auf den Transkriptionen ist es möglich eine Vielzahl von Auszeichnungen und Annotationen innerhalb des Textes, aber auch darüber hinaus für Einzeldokumente und ganze Dokumentenbestände anzulegen. Neben der Anreicherung der Dokumente mit Metadaten (Identifikation von Personen, Orten und Sachwörtern) ist somit auch die Möglichkeit der Herstellung von Bestandsbeschreibungen und der Hinterlegung von Transkriptions- und Editionsvorschriften gegeben.

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. Ausgehend davon können komplexe digitale Editionen erstellt werden, die jedoch im Unterschied zu den meisten herkömmlichen Editionen eine enge Verzahnung mit den verwendeten Bilddateien aufweisen. Dadurch werden Editionen ermöglicht, die den transkribierten Text in der Zusammenschau mit der faksimilierten Vorlage sichtbar machen. Daneben sind auch Ausgaben als Druckdaten (PDF) oder zur Weiterbearbeitung für Textverarbeitungsprogramme (DOCX) 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 ist für unterschiedliche Gruppen konzipiert. Einerseits für GeisteswissenschaftlerINNEN, 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 Aufbereitung 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 GeisteswissenschaftlerINNEn und ComputerspezialistINNen 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 ComputerwissenschaftlerINNEN 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 GeisteswissenschaftlerINNEN als auch an ComputerwissenschaftlerINNEN, wobei vorwiegend die Tools und Möglichkeiten von Transkribus präsentiert werden.

Zwei zentrale Forschungsaspekte aus READ werden im Rahmen des Workshops durch Experten vorgestellt:

Einerseits das technische Verfahren des Automatic Text Recognition mit rekurrenten neuronalen Netzen (Leifert et al. 2016). Dabei wird kurz in die Trainings- und Auswertungsmechanismen mit neuronalen Netzen eingeführt und Möglichkeiten der Auswertung demonstriert.

Andererseits wird die Erkennung von komplexen Layouts, insbesondere Tabellen, erklärt und neueste technische Lösungen vorgestellt.

Programm/Ablauf des Workshops

- *Begrüßung und Informationen zum Projekt READ* (Tobias Hodel, Zürich): 20‘
  Überblick über Ziele und Fortschritte im Rahmen des von der EU geförderten Projekts.
- *Machine Learning und automatisierte Text Erkennung* (Tobias Strauß, Rostock): 30‘
  Einführung und Erklärung zum Einsatz neuronaler Netze bei der Texterkennung
- *Einführung in Transkribus* 30‘
  Aufbau und Funktionieren des Programms, Demonstration des Gebrauchs anhand von Beispielen. Aufzeigen der Möglichkeiten zum Einsatz der Automatisierungen.
- *Selbstständiges Arbeiten der Teilnehmenden mit Transkribus:* 90‘
- Die Möglichkeiten und Grenzen von Transkribus sollen von den Teilnehmenden (falls gewünscht mit eigenen Dokumenten) selbst ausgetestet werden.
- *Layout Analyse: Tabellen und andere schwierige Formen* (Markus Diem, Wien): 30‘
  Ein über Transkribus hinausgehender Teil des Projekts beschäftigt sich mit *computer vision*. Ziel ist es, auch komplexe Strukturen korrekt als Layout zu erkennen, um die automatisierte Texterkennung überhaupt zu ermöglichen. Tabellen gehören in dem Bereich zu den schwierigsten Formen der Texterkennung.
- *Diskussion über Vor- und Nachteile der Software:* 30‘
  Inklusive Evaluation des Tools und der Veranstaltung. Feedbacks werden eingeholt, zur Verbesserung der Software und Webtools (usability, Umfang und Leistung der Automatisierungen etc.).
- Nach Interesse der Teilnehmenden werden während des Workshops Kurzinputs zu folgenden Themen angeboten:
  1. Matching von Text und Bild (bspw. aus bestehenden Transkriptionen),
  2. Transkribus Learn (e-Learningumgebung),
  3. Crowdsourcing-Infrastruktur,
  4. ScanTent und DocScan (Fotografieren eigener Dokumente mit Android App).

Workflow in Transkribus

Um Texte zu transkribieren oder zu edieren, müssen digitale Bilder hochgeladen und danach mit Layouterkennungswerkzeugen bearbeitet werden. Die Analyse des Layouts kann automatisiert geschehen, wobei die manuelle Kontrolle und falls nötig die Nachbearbeitung im Moment noch sinnvoll ist.

Dokumente können entweder automatisch mit bereits bestehenden ATR-Modellen (Automatic Text Recognition) erkannt werden oder die Transkription erfolgt händisch und kann danach zum Training neuer Modelle genutzt werden. Insbesondere für die Bearbeitung großer Dokumentenkorpora, die in ähnlichen Handschriften verfasst wurden, lassen sich bereits heute Effizienzgewinne und Vereinfachungen erzielen.

Aufbauend auf den Transkriptionen ist es möglich eine Vielzahl von Auszeichnungen und Annotationen innerhalb des Textes, aber auch darüber hinaus für Einzeldokumente und ganze Dokumentenbestände anzulegen. Neben der Anreicherung der Dokumente mit Metadaten (Identifikation von Personen, Orten und Sachwörtern) ist somit auch die Möglichkeit der Herstellung von Bestandsbeschreibungen und der Hinterlegung von Transkriptions- und Editionsvorschriften gegeben.

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. Ausgehend davon können komplexe digitale Editionen erstellt werden, die jedoch im Unterschied zu den meisten herkömmlichen Editionen eine enge Verzahnung mit den verwendeten Bilddateien aufweisen. Dadurch werden Editionen ermöglicht, die den transkribierten Text in der Zusammenschau mit der faksimilierten Vorlage sichtbar machen. Daneben sind auch Ausgaben als Druckdaten (PDF) oder zur Weiterbearbeitung für Textverarbeitungsprogramme (DOCX) 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 ist für unterschiedliche Gruppen konzipiert. Einerseits für GeisteswissenschaftlerINNEN, 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 Aufbereitung 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 GeisteswissenschaftlerINNEn und ComputerspezialistINNen 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 ComputerwissenschaftlerINNEN 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 GeisteswissenschaftlerINNEN als auch an ComputerwissenschaftlerINNEN, wobei vorwiegend die Tools und Möglichkeiten von Transkribus präsentiert werden.

Zwei zentrale Forschungsaspekte aus READ werden im Rahmen des Workshops durch Experten vorgestellt:

Einerseits das technische Verfahren des Automatic Text Recognition mit rekurrenten neuronalen Netzen (Leifert et al. 2016). Dabei wird kurz in die Trainings- und Auswertungsmechanismen mit neuronalen Netzen eingeführt und Möglichkeiten der Auswertung demonstriert.

Andererseits wird die Erkennung von komplexen Layouts, insbesondere Tabellen, erklärt und neueste technische Lösungen vorgestellt.

Programm/Ablauf des Workshops

Begrüßung und Informationen zum Projekt READ (Tobias Hodel, Zürich): 20‘

Überblick über Ziele und Fortschritte im Rahmen des von der EU geförderten Projekts. *Machine Learning und automatisierte Text Erkennung* (Tobias Strauß, Rostock): 30‘

Einführung und Erklärung zum Einsatz neuronaler Netze bei der Texterkennung *Einführung in Transkribus* 30‘

Aufbau und Funktionieren des Programms, Demonstration des Gebrauchs anhand von Beispielen. Aufzeigen der Möglichkeiten zum Einsatz der Automatisierungen. *Selbstständiges Arbeiten der Teilnehmenden mit Transkribus:* 90‘

Die Möglichkeiten und Grenzen von Transkribus sollen von den Teilnehmenden (falls gewünscht mit eigenen Dokumenten) selbst ausgetestet werden.

Layout Analyse: Tabellen und andere schwierige Formen (Markus Diem, Wien): 30‘

Ein über Transkribus hinausgehender Teil des Projekts beschäftigt sich mit *computer vision*. Ziel ist es, auch komplexe Strukturen korrekt als Layout zu erkennen, um die automatisierte Texterkennung überhaupt zu

ermöglichen. Tabellen gehören in dem Bereich zu den schwierigsten Formen der Texterkennung.

Diskussion über Vor- und Nachteile der Software: 30‘ Inklusive Evaluation des Tools und der Veranstaltung. Feedbacks werden eingeholt, zur Verbesserung der Software und Webtools (usability, Umfang und Leistung der Automatisierungen etc.).

Nach Interesse der Teilnehmenden werden während des Workshops Kurzinputs zu folgenden Themen angeboten:
- Matching von Text und Bild (bspw. aus bestehenden Transkriptionen),
- Transkribus Learn (e-Learningumgebung),
- Crowdsourcing-Infrastruktur,
- ScanTent und DocScan (Fotografieren eigener Dokumente mit Android App).

Während des gesamten Workshops stehen drei 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: 30-40 Personen (auch abhängig von der Raumgrösse).
Benötigte technische Ausstattung: Allgemein: Beamer, evtl. Whiteboard.
Teilnehmende: Eigener Rechner (wenn möglich Installation von Transkribus; Hilfe zur Installation von Transkribus wird 15 Minuten vor der Veranstaltung angeboten)

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

Kontaktdaten aller Beitragenden (inkl. Forschungsinteressen)

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 Textrecognition; eArchiving; Information Retrieval).

Tobias Strauß, Institut für Mathematik, Ulmenstraße 69, Universität Rostock, 18051 Rostock, Deutschland; tobias.strauss@uni-rostock.de; (Deep Learning, Information Retrieval und Natural Language Processing).

Bibliographie

Leifert, G., Strauß, T., Grüning, T., Wustlich, W., Labahn, R., 2016. Cells in Multidimensional Recurrent Neural Networks. Journal of Machine Learning Research 17, 1-37.

Leifert, G., Strauß, T., Grüning, T., Wustlich, W., Labahn, R., 2016. Cells in Multidimensional Recurrent Neural Networks. Journal of Machine Learning Research 17, 1-37.