Automatische Texterkennung von Handschriften und historischen Drucken

Qualität und Normierung von Ground-Truth-Daten in der Praxis

Huff, Dorothee
dorothee.huff[at]uni-tuebingen.de
Universitätsbibliothek Tübingen, Deutschland
ORCID-iD: 0000-0003-0866-9967

Will, Larissa
larissa.will[at]uni-mannheim.de
Universitätsbibliothek Mannheim, Deutschland
ORCID-iD: 0009-0004-6220-8939

Stöbener, Kristina
kristina.stoebener[at]uni-tuebingen.de
Universitätsbibliothek Tübingen, Deutschland
ORCID-iD: 0000-0002-3299-974X

Zusammenfassung: Automatische Texterkennung (OCR) übersetzt textliche Bildinhalte in digitale Textformate. Auf diese Weise werden der Zugang zu historischen Drucken und Handschriften erhöht und neue Forschungsfragen an das Material ermöglicht. Vor der wissenschaftlichen Auswertung der Daten gilt es jedoch, sich über Aspekte wie Qualität und Normierung der Ground-Truth-Daten und des erzeugten Outputs bewusst zu werden, diese zu hinterfragen und bei der Nachnutzung der Daten in Betracht zu ziehen. Anhand von Beispielen sollen unterschiedliche Vorgehensweisen bei der Erzeugung von Ground-Truth-Daten sowie Ergebnisse der jeweiligen Modelltrainings vorgestellt und problematisiert werden.

Automatische Texterkennung (OCR) übersetzt textliche Bildinhalte in digitale Textformate. Auf diese Weise werden der Zugang zu historischen Drucken und Handschriften erhöht und neue Forschungsfragen an das Material ermöglicht. Erkannte Texte können durchsucht, kopiert, bearbeitet und für eine Extraktion von Forschungsdaten verwendet werden. Große Mengen an Text können bei geringem Ressourceneinsatz – im Gegensatz zur manuellen Transkription – erzeugt und verarbeitet werden. Auch paläographische Kenntnisse sind auf diesem Weg keine zwingende Voraussetzung für die Arbeit mit primärem Quellenmaterial.

Während die automatische Texterkennung von historischen Dokumenten mit klassischer OCR kaum möglich war, wurden mithilfe von Machine Learning in den letzten Jahren große Fortschritte gemacht.[1] Moderne Texterkennungsmodelle sind neuronale Netze, die anhand von korrigierten Transkriptionen, sogenannten Ground-Truth-Daten, trainiert werden. Während der Begriff „Ground-Truth“ auf den Ausschließlichkeitsanspruch nur einer möglichen richtigen Lösung hinzudeuten scheint, können die zugrunde liegenden historischen Schriftzeichen bei der Transkription jedoch oftmals unterschiedlich interpretiert und wiedergegeben sowie sogar je nach Zielsetzung bewusst an individuelle Standards angepasst werden. So kann zwar die Weiterverarbeitung der Daten in spezifischen Kontexten wie z. B. als Editionsgrundlage, für linguistische Untersuchungen oder die Einbringung in Datenbanken erleichtert werden, allerdings entsteht auf diese Weise ein Spannungsverhältnis eines solchermaßen passgenau personalisierten Datensatzes zu einer allgemeinen Nachnutzbarkeit der Daten. Zudem fehlt gerade bei historischen Dokumenten oftmals eine Normierung bezüglich der Wiedergabe von historischen Schriftzeichen mit einem Codepoint, damit diese einheitlich von Maschinen verarbeitet werden können. Eine Lösung zumindest für den Bereich historischer Druckwerke sind die Ground-Truth-Richtlinien der koordinierten Förderinitiative OCR-D, die normierte Handlungsempfehlungen für drei Transkriptionslevel bieten.[2]

Die Erzeugung von Ground-Truth-Daten für das Training von Texterkennungsmodellen erfolgt an den Universitätsbibliotheken Tübingen und Mannheim zeichennah nach Level 2 der OCR-D-Transkriptionsrichtlinien, um eine möglichst breite Nachnutzbarkeit der Daten zu gewährleisten und Transformationsmöglichkeiten offenzuhalten. Da der Standardisierungsprozess noch nicht abgeschlossen ist und nicht alle Fälle abgedeckt werden, ist oftmals zwangsweise doch wieder ein zumindest in Teilen individualisiertes Vorgehen notwendig. In der Praxis ist daher zu überlegen, inwieweit dies die Weiterverarbeitung der Daten beeinflusst, wie bei Veröffentlichung der Daten mit Abweichungen umzugehen ist und wie konvertibel die Daten sind. Ist der diplomatische Ansatz jedoch überhaupt und wenn ja, in welchen Kontexten sinnvoll? So gewährleistet die alternative Herangehensweise einer Normalisierung der Daten in der Regel eine bessere Lesbarkeit und Durchsuchbarkeit. Es soll beispielhaft analysiert werden, wie sich unterschiedliche Transkriptionsrichtlinien auf die Erzeugung standardisierter Daten und die Zeichenfehlerrate der Texterkennungsmodelle auswirken.

[1] Hodel 2023, 154–157.
[2] OCR-D, n.d.

Die für die Erstellung der Ground-Truth-Daten getroffenen Entscheidungen haben jedoch nicht nur einen Einfluss auf die Nachnutzung dieser Daten in weiteren Kontexten, wie z. B. bei der Zusammenstellung verschiedener Ground-Truth-Datensätze für Modelltrainings, wo uneinheitliche Transkriptionsrichtlinien zu Problemen führen können, sondern bestimmen im nächsten Schritt auch den Output der auf Grundlage dieser Daten trainierten Texterkennungsmodelle. Unter Umständen können zwar regelhaft verwendete Zeichen sowohl in den Ground-Truth-Daten wie auch in der automatisch erzeugten Transkription im Nachhinein ersetzt und die Daten so für andere Nutzungskontexte angepasst werden, jedoch hängt das Ergebnis der automatischen Transkription grundsätzlich von den bei der Erzeugung des Trainingsmaterials getroffenen Entscheidungen und der Qualität desselben ab. Egal wie gut letztere ist, muss beachtet werden, dass mit den aktuellen technischen Möglichkeiten in der Regel kein hundertprozentig korrektes Ergebnis erzielt werden kann, und überlegt werden, wie damit umzugehen ist.[3]

Beim Einsatz von Texterkennungssoftware auf historische Dokumente gilt es somit, sich verschiedener Fragestellungen bei der Datenerzeugung und -nachnutzung bewusst zu werden. Das 2019 im Rahmen des Projekts *OCR-BW* als Service der Universitätsbibliotheken Tübingen und Mannheim eingerichtete *Kompetenzzentrum „Volltexterkennung von handschriftlichen und gedruckten Werken“* betreut Wissenschaftlerinnen und Wissenschaftler sowie Bibliotheken und Archive in Baden-Württemberg bei der Anwendung von automatischer Texterkennungs- und Transkriptionssoftware.[4] Anhand eigener Textkorpora aus Beständen der UB Tübingen, wie z. B. Expeditionstagebüchern, juristischen Konsilien und mittelalterlichen Handschriften, wie auch bei der Unterstützung von wissenschaftlichen Projekten aus verschiedenen Fachdisziplinen werden die Transkriptionsplattformen *Transkribus*[5] und *eScriptorium*[6] für die Erzeugung von automatischen Volltexten für Handschriften und Drucke systematisch getestet und eingesetzt.[7] Anhand von Beispielen sollen unterschiedliche Vorgehensweisen bei der Erzeugung von Ground-Truth-Daten sowie Ergebnisse der jeweiligen Modelltrainings vorgestellt und problematisiert werden.

[3] Neudecker et al. 2021, 153–157.
[4] OCR-BW, n.d.
[5] Transkribus, n.d.
[6] Scripta/eScriptorium, n.d.
[7] Huff and Stöbener 2022.

Bibliografie

Hodel, Tobias. “Konsequenzen der Handschriftenerkennung und des maschinellen Lernens für die Geschichtswissenschaft. Anwendung, Einordnung und Methodenkritik.” *Historische Zeitschrift* 316, no. 1 (February 2023): 151–180. https://doi.org/10.1515/hzhz-2023-0006.

Huff, Dorothee, and Kristina Stöbener. “Projekt OCR-BW: Automatische Texterkennung von Handschriften.” *O-Bib. Das Offene Bibliotheksjournal / Herausgeber VDB* 9, no. 4 (2022): 1–19. https://doi.org/10.5282/o-bib/5885.

Neudecker, Clemens, Karolina Zaczynska, Konstantin Baierer, Georg Rehm, Mike Gerber and Julián Moreno Schneider. “Methoden und Metriken zur Messung von OCR-Qualität für die Kuratierung von Daten und Metadaten.” In *Qualität in der Inhaltserschließung*, edited by Michael Franke-Maier, Anna Kasprzik, Andreas Ledl and Hans Schürmann, 137–166. Berlin, Boston: De Gruyter Saur, 2021. https://doi.org/10.1515/9783110691597-009.

OCR-BW. “OCR-BW. Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen.” Accessed July 21, 2023. https://ocr-bw.bib.uni-mannheim.de/.

OCR-D. “Die Ground Truth Richtlinien.” Accessed July 21, 2023. https://ocr-d.de/de/gt-guidelines/trans/.

Scripta/eScriptorium. “scriptorium.” Accessed July 21, 2023. https://gitlab.com/scripta/escriptorium/.

Transkribus. “READ-COOP – Transkribus.” Accessed July 21, 2023. https://readcoop.eu/transkribus/.