HTR

Mühlberger, Günter; guenter.muehlberger@uibk.ac.at

Die automatisierte Erkennung von Handschrift (*Handwritten Text Recognition*) beruht in *Transkribus* auf exakt demselben Verfahren bzw. derselben *Engine*, die auch für die Druckschriftenerkennung (OCR) angewendet wird. Allerdings müssen die neuronalen Netze mit einer Reihe zusätzlicher Herausforderungen zurechtkommen, da der Standardisierungsgrad bei Handschriften insgesamt wesentlich geringer ist als bei Druckschrift. Konkret bedeutet dies, dass wesentlich mehr Trainingsdaten notwendig sind, um diese Aufgabe bewältigen zu können. Das macht sich insbesondere bei großen Modellen bemerkbar, die zum Beispiel für eine Epoche von 100 oder mehr Jahren gute Ergebnisse erzielen sollen.

Der einfachste Fall, der jedoch bei der Erstellung von digitalen Editionen häufig auftreten wird, ist gegeben, wenn ein Modell für einen einzelnen Schreiber trainiert werden soll. Hier reichen schon relativ wenige Seiten aus, um gute Ergebnisse erzielen zu können. Als Beispiel führen wir die Tagebücher von Andreas Okopenko an, die im Rahmen des KONDE-Projekts in *Transkribus* trainiert wurden. Hier lässt sich auch gut der Fortschritt der letzten Jahre dokumentieren. Das erste Modell, das im Frühjahr 2018 erzeugt wurde, weist eine Fehlerquote von 10,17 Prozent am Validierungsset auf. Mit der weiterentwickelten Engine hingegen wird auf den identischen Trainingsdaten eine Fehlerquote von 3,61 Prozent erreicht. Das Trainingsset besitzt in beiden Fällen 20.782 Wörter, geht man von 200 Wörtern pro Seite aus, dann liegen also nicht mehr als ca. 100 Seiten Trainingsmaterial zugrunde.

Abbildung: Beispiel Texterkennung - Andreas Okopenko: Fehlerquote auf dieser Seite: 1,76

Eines der größten Modelle für historische Handschriften in *Transkribus* wurde vom Nationalarchiv der Niederlande zusammen mit dem Stadtarchiv Amsterdam erstellt. Das Modell umfasst ca. 7.000 Seiten bzw. 1.384.893 Wörter und erzielt auf dem Validierungsset 5,67 Prozent. Die zugrundeliegenden Trainingsdaten wurden auf Basis einer Zufallsstichprobe aus mehreren Millionen Seiten des 18. Jahrhunderts ausgewählt. Das Modell enthält hunderte unterschiedliche Schreiber und kann daher mit einer Vielzahl unterschiedlicher Schreibstile umgehen.

Abbildung: Beispiel Texterkennung - Niederländisches Dokument

Ganz ähnliche Ergebnisse können auch mit den Kurrentmodellen in *Transkribus* erzielt werden, die auf ähnlichen Datenmengen beruhen. Hier wurden mehrere tausend Seiten Kurrentschrift aus dem 17. bis 20. Jahrhundert zugrun-

de gelegt. Der Schwerpunkt liegt allerdings auf dem späten 19. Jahrhundert. Die Modelle sind in *Transkribus* frei verfügbar.

Zusammenfassend lässt sich sagen, dass mit dem Einsatz moderner Methoden der Texterkennung bei historischen Druckschriften nahezu fehlerlose Transkriptionen erzielt werden können. Bei historischen Handschriften sind die Ergebnisse noch deutlich fehlerhafter, trotzdem lassen sich mit überschaubarem Aufwand auch für Handschriften Modelle trainieren, die sowohl die Transkription beschleunigen, als auch die Durchsuchbarkeit großer Dokumentenmengen ermöglichen.

Literatur:

- Mühlberger, Günter; Colutto, Sebastian; Kahle, Philipp: Handwritten Text Recognition (HTR) of Historical Documents as a Shared Task for Archivists, Computer Scientists and Humanities Scholars: The Model of a Transcription amp; Recognition Platform (TRP).
- Strauß, Tobias; Weidemann, Max; Michael, Johannes; Leifert, Gundram; Grüning, Tobias; Labahn, Roger: System Description of CITlab's Recognition amp; Retrieval Engine for ICDAR2017 Competition on Information Extraction in Historical Handwritten Records. In: arXiv:1804.09943 [cs]: 2018.
- READ-COOP SCE: Public Models in Transkribus: 2020. URL: `https://readcoop.eu/transkribus/public-models/`.

Software:

HTR+, PyLaia, Tesseract, The OCRopus OCR System and Related Software, SimpleHTR, Transkribus

Verweise:

OCR, Diplomatische Transkription, Transkription, Transkriptionswerkzeuge

Projekte:

Tagebücher Andreas Okopenko, Noscemus, Newseye, Transkribus

Themen:

Digitalisierung

Zitiervorschlag:

Mühlberger, Günter. 2021. HTR. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". URL: https://gams.uni-graz.at/o:konde.224