OCR

Fritze, Christiane; christiane.fritze@onb.ac.at / Mühlberger, Günter; guenter.muehlberger@uibk.ac.at

OCR steht für *Optical Character Recognition*, also für 'optische Zeichenerkennung'. Eine OCR-Software liest ein Bilddigitalisat ein und erkennt darauf vorhandenen Text, sodass dieser nach der OCR-Erkennung durchsuchbar ist (STRG F). Bevor die OCR-Software mit dem Erkennungsdurchlauf beginnt, wird das zu analysierende Bilddigitalisat binarisiert, d. h. aus einem Farbscan wird ein bitonales Bild erstellt. Anschließend wird das Layout des Bilddigitalisats erfasst und in Erkennungszonen eingeteilt. Für die Erkennung werden die einzelnen Pixelcluster mit in der Software verankerten Pixelmustern abgeglichen und der wahrscheinlichste Wert ausgewählt.

Der Leistungsumfang von Softwareangeboten ist unterschiedlich: Je nach Software und Ausgabeformat kann die Erkennungswahrscheinlichkeit in den Metadaten vermerkt werden. Je nach Software kann für die Erkennung ein Wörterbuch für bestimmte Sprachen oder diachrone Sprachstufen hinterlegt sein. Je nach Bedarf können die zu analysierenden Zonen vorab markiert werden bzw. ist eine händische Fehlerkorrektur des erkannten Textes möglich.

Für die Weiterverarbeitung stehen verschiedene Ausgabeformate bereit: einfacher Text, pageXML, das XML-basierte Format ALTO oder hOCR mit Layoutinformationen.

Optical Character Recognition gehört zu den bekanntesten Anwendungen der Mustererkennung und wurde bereits in den 70er-Jahren von Raymond Kurzweil als Standardverfahren entwickelt. Zuerst nur für spezielle Schriften geeignet, erweiterten sich die Einsatzmöglichkeiten Schritt für Schritt. In den späten 90er-Jahren erreichte die Erkennungsgenauigkeit einen ersten Höhepunkt: Geschäftsdokumente wie Briefe, Rechnungen, aber auch Bücher und moderne Zeitungen konnten nun zuverlässig erkannt werden. Doch Schriften aus dem 15. bis zum 19. Jahrhundert konnten trotz intensiver Bemühungen auch weiterhin nicht zufriedenstellend gelesen werden. Ein bekanntes Beispiel hierfür sind die anfangs von *Google* gelieferten Volltexte im Zusammenhang mit dem *Google Books*-Projekt.

Die Revolution in der Erkennung von herausfordernden, nicht-standardisierten Schriften fand erst vor wenigen Jahren statt und beruht auf neuen Methoden des Maschinenlernens, die zuvor schon bei der automatisierten Sprach- und Handschriftenerkennung entwickelt wurden. Statt eine Zeile in Wörter oder gar einzelne Buchstaben zu zerlegen, wie dies bei traditioneller OCR der Fall ist, wird nunmehr das Bild einer vollständigen Zeile einem neuronalen Netz zusammen mit dem dazugehörigen Text als Trainingsmaterial vorgesetzt. Das Netz lernt dann selbständig den Zusammenhang zwischen Bild und Text. Wie die nachfolgenden Beispiele zeigen werden, können mit der nunmehr dominierenden Methode der Texterkennung sowohl historische Bücher, als auch beliebige

Handschriften (HTR) mit einer erstaunlichen Genauigkeit erkannt werden.

Für unsere Beispiele verwenden wir Zahlen, die wir im Rahmen der Arbeit mit der Transkribus-Plattform gemacht haben. Dort sind zwei OCR-*Engines* im Einsatz, einmal die vom CITlab-Team der Universität Rostock entwickelte HTR+ und zum andern die von der PRHLT-Gruppe der Universität Valencia entwickelte *PyLaia Engine*. Letztere ist auch als Open Source-Software verfügbar. Für unsere Experimente haben wir die HTR+-*Engine* verwendet. Daneben gibt es noch eine Reihe anderer OCR-Anwendungen, die bekanntesten darunter sind wahrscheinlich *Tesseract* von *Google* oder *Ocropy* von Thomas Breuel, zusätzlich haben auch *Google, Amazon, Microsoft* und *Facebook* jeweils eigene OCR-*Engines* entwickelt und bieten diese über ihre Plattformen an.

Das erste Modell, das wir kurz vorstellen wollen, wurde im Rahmen des *NewsEye*-Projekts erstellt und 2019 auf der *Transkribus*-Plattform veröffentlicht. Ein Update ist in Vorbereitung. Das Modell beruht auf Trainingsdaten österreichischer Zeitungen des späten 19. und frühen 20. Jahrhunderts im Umfang von 442.141 Wörtern. Das Ergebnis am Validierungsset beträgt 1,66 Prozent *Character Error Rate*. Diese Messung beinhaltet auch Satzzeichen, Zahlen in Tabellen, Werbeeinschaltungen und ähnliches, liefert also ein eher pessimistisches Ergebnis. Für laufenden Text werden hingegen deutlich bessere Fehlerraten von unter 1leistungsfähig das Netz ist, kann im folgenden Beispiel gezeigt werden. Es handelt sich um einen Ausschnitt aus der *Rheinischen Volksstimme* aus dem frühen 20. Jahrhundert. Die Bildqualität ist aufgrund des schlechten Drucks (durchscheinende Buchstaben) sowie der Tatsache, dass hier ein Mikrofilm der Digitalisierung zugrunde lag, stark eingeschränkt.

Abbildung: Beispiel Texterkennung - Vergleich Abbyy FineReader - Google OCR - Transrkibus

Noch bessere Ergebnisse lassen sich mit dem Modell *Noscemus* erzielen, das im Rahmen eines ERC-Grants am Institut für Neulatein der Universität Innsbruck erstellt und in *Transkribus* veröffentlicht wurde. Es zeigt eine durchschnittliche Fehlerquote am Validierungsset von 0,86 Prozent CER auf. Hier liegen dem Modell neulateinische Schriften des 16. bis 18. Jahrhunderts im Umfang von 307.329 Wörtern zugrunde. Zudem wurde diesem Netz auch beigebracht, die typischen Abkürzungen wie sie im Neulateinischen gebräuchlich sind, gleich bei der Erkennung aufzulösen.

Abbildung: Beispiel Texterkennung - Neulateinisches Dokument

Literatur:

- READ-COOP SCE: Public Models in Transkribus: 2020. URL: `https://readcoop.eu/transkribus/public-models/`.

Software:

Abby Finereader, Cuneiform, Tesseract, The OCRopus OCR System and Related Software, Transkribus, HTR+, PyLaia, OCR4all, OCR-d, Virtual Transcription Laboratory

Projekte:

Google Books, Transkribus

Verweise:

Digitalisierung, HTR, Transkription, Transkriptionswerkzeuge

Themen:

Digitalisierung

Lexika

- Edlex: Editionslexikon
- Lexicon of Scholarly Editing

Zitiervorschlag:

Fritze, Christiane; Mühlberger, Günter. 2021. OCR. 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.149