Zur OCR frühneuzeitlicher Drucke am Beispiel des RIDGES-Korpus von Kräutertexten

Uwe Springmann (LMU München und Humboldt-Universität zu Berlin)
& Anke Lüdeling (Humboldt-Universität zu Berlin)

Unser Vortrag stellt eine Methode zur optischen Zeichenerkennung (OCR) von frühen Drucken vor, die deutlich bessere Resultate zeigt als vorherige Methoden. Mithilfe des Verfahrens können leichter und schneller Korpora mit frühen Texten erstellt werden, die dann nur noch nachkorrigiert werden müssen. Mit dem Aufbau solcher Korpora aus frühneuzeitlichen Drucken werden Basisressourcen für alle darauf aufbauenden Forschungen sprachlicher, historischer und kulturgeschichtlicher Art in den Digital Humanities bereitgestellt. Wir exemplifizieren unsere Methode mit Daten aus dem RIDGES-Korpus[1], einem diachronen Korpus, das deutschsprachige Kräutertexte enthält, die zwischen 1487 und 1870 entstanden sind.

Mangels maschineller Unterstützung ist die Erstellung eines solchen Korpus aufwändig und vom Finden korpusrelevanter gut lesbarer Vorlagen über die Einweisung von Hilfskräften in die Transkription ungewohnter Zeichen und paläographischer Konventionen und einer für die Korrektur der Transkription notwendigen breiten Sprach- und Sachkenntnis geprägt. Eine historische Orthographie und der unvermittelte Wechsel von deutschem Fraktur-Text zu lateinischen Zitaten in Antiqua sowie griechischen Wörtern erschweren die Erstellung der Transkription zusätzlich. Insbesondere die frühen Drucke (Wiegendrucke, aber auch noch Drucke aus dem 17. Jahrhundert) sind hier schwierig.

Der Traum von einer automatischen Unterstützung bei der Konvertierung früher Drucke durch allgemein zugängliche Methoden einer OCR, die ein entsprechendes Training der Erkennungsroutinen auf die verwendeten Schriften sowie die Eigentümlichkeiten des Druckbildes voraussetzen, ließ sich bisher angesichts proprietärer, einem umfangreichen Training für Außenstehende nicht zugänglicher Industrieprodukte (z.B. Abbyy Finereader[2]) bzw. zwar quelloffener und grundsätzlich trainierbarer, aber an der gestellten Aufgabe scheiternder Software (z.B. Tesseract[3]) nicht verwirklichen. Neben diesen grundsätzlichen Mängeln stand einem solchen Ansatz bisher auch der Umstand entgegen, dass ein Training eine systematisch erstellte diplomatische, d.h. am Druckbild orientierte und nicht-normalisierende Transkription von Texten voraussetzt.

Im Jahr 2013 wurden die bei Mustererkennungsaufgaben sehr erfolgreichen rekurrenten neuronalen Netzwerke mit langem Kurzzeitgedächtnis (LSTM: long short-term memory; Hochreiter & Schmidhuber 1997) durch Thomas Breuel erstmals in die OCR eingeführt und in das quelloffene, schon länger bestehende System OCRopus (Version 0.7)[4] integriert (Breuel et al. 2013). Das RIDGES-Korpus enthält Textausschnitte aus vielen Kräuterbüchern. Diese Ausschnitte (meist ca. 30 Textseiten) wurden eng diplomatisch transkribiert.[5] Das Training dieses Systems mit Hilfe der diplomatischen Transkription ("ground truth") zeigt Ergebnisse, die bei jedem der vorliegenden Texte eine Rate korrekt erkannter Zeichen (einschließlich Ligaturen, Diakritika

---

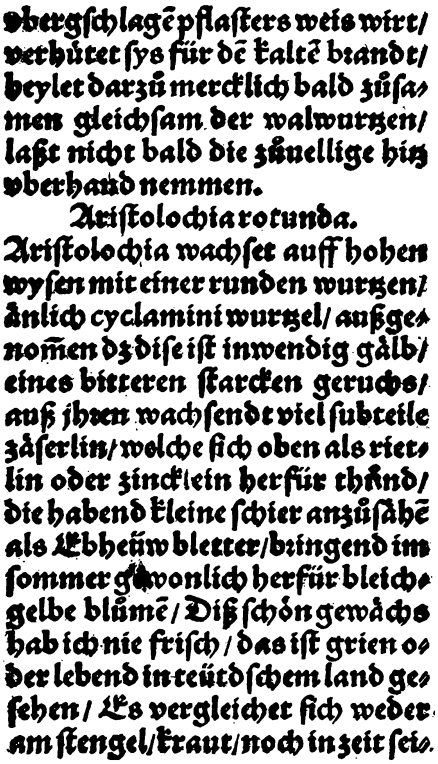
1 Ridges steht für Register in Diachronic German Science; Ziel des Ridges-Projekts ist die qualitative und quantitative Analyse der Entstehung eines deutschsprachigen wissenschaftlichen Registers. Dazu gibt es viel Literatur (so z. B. Klein 2011 oder Habermann 2003), die sich bisher aber auf nicht digital vorliegende Texte stützen musste und daher kaum für statistische Registeranalysen ausgewertet werden konnte. Das Korpus ist unter der CC-BY-Lizenz verfügbar unter http://korpling.german.hu-berlin.de/ridges/index_de.html. Das Korpus ist tief annotiert und wächst ständig.

2 http://www.abbyy.de/

3 http://code.google.com/p/tesseract-ocr/

4 http://www.ocropus.com

und Leerzeichen) von über 96% selbst ohne Verwendung von Sprachmodellen und Nachkorrekturen ergibt, während bisherige Versuche mit kommerziell erhältlicher Software bzw. Tesseract, an denen ein Autor seit Jahren beteiligt ist, kaum an die Grenze von 90% heranreichen (Springmann et al. 2013).[6]



Vbergſchlagēpflaſters weis wirt / verhůtet ſys für dē kaltē brandt / heylet darzů mercklich bald zůſa= inen gleichſam der walturtzen / laſzt nicht bald die zůuellige hitz tberhaud nemmen.
Ariſtolochia rotunda.
Ariſtolochia wachſet auff hohen wyſen mit einer runden wurtzen / ånlich cyclamini wurtzel / auſzge= nomēn dz diſe iſt inwendig gålb / eines bitteren ſtarcken geruchs / auſz jhren wachſendt viel ſubteile zåſerlin / welche ſich oben als riet= lin oder zincklein herfür thůnd / die habend kleine ſchier anzůſåhē als Ebheüw bletter / bringend im ſommer gvonlich herfür bleich= gelbe blůmē / Diſz ſchön gewåchs hab ich nie friſch / das iſt grien o= der lebend in teütdſchem land ge= ſehen / Es vergleichet ſich weder am ſtengel / kraut / noch in zeit ſei=

Adam von Bodenstein (1557): Wie sich meniglich … . Unkorrigierter OCR–Output einer vorher nicht gesehenen Seite nach Training auf 49.000 zufällig ausgewählten Textzeilen (Bild + zugeordnete ground truth) aus einer Traningsmenge von 34 diplomatisch transkribierten Seiten. Die OCR zeigt 7 verbleibende Fehler auf dieser Seite (das entspricht der durchschnittlichen Zeichenerkennungsrate auf einer Testmenge von Seiten von 99,0%).

Der Grund für diese hochgenaue Erkennungsrate liegt darin, dass OCRopus im Gegensatz zu bisherigen Methoden keine Erkennung auf Zeichenbasis über ein Template-Matching-Verfahren durchführt, bei dem ein errechnetes „mittleres“ Zeichen (das Template) auf Übereinstimmung mit einem zu erkennenden Zeichen überprüft wird, sondern jede Druckzeile durch Zerlegung in bis zu 1000 vertikale

5 Die Transkription wurde von Studierenden in verschiedenen Seminaren begonnen und später korrigiert. Daneben gibt es zwei Normalisierungsebenen und verschiedene Annotationsebenen.

6 Lediglich für die kommerzielle Software B.I.T. Alpha wurden ähnlich gute Ergebnisse für Drucke des 16. und 17. Jahrhunderts berichtet, wobei die erreichbare Genauigkeit von einem für Außenstehende kaum nachzuvollziehenden In-House-Training des kommerziellen Anbieter abzuhängen scheint (Stäcker in Federbusch et al. 2013).

Streifen schneidet, so dass jeder Buchstabe und jeder Wortzwischenraum in bis zu 30 Streifen zerlegt wird. Für jeden Streifen werden im Laufe des Trainings über den Vergleich von gedruckter Zeile mit ihrer zugeordneten Transkription die Parameter des neuronalen Netzes so eingestellt, dass mit hoher Wahrscheinlichkeit der richtige Buchstabe ausgegeben wird. Die Übersegmentierung der Buchstaben führt zu einer höheren Auflösung bei der Erkennung, so dass auch zwischen ähnlichen Zeichen wie langem s (ſ) und f problemlos unterschieden werden kann.

Die Aussicht, dass sich nunmehr jeder Interessierte Texte in hoher Genauigkeit in elektronischer Form verschaffen kann, selbst wenn die zugrundeliegenden Drucke aus früheren Jahrhunderten stammen, wird anhand unserer Erfahrungen mit dem RIDGES-Korpus hinsichtlich ihrer Voraussetzungen und des damit verbundenen Aufwandes kritisch beleuchtet. Dabei werden sowohl die Rolle des Trainings als auch der Nachkorrektur sowie die Stellung der OCR im gesamten Prozess der Korpuserstellung diskutiert. Die verwendeten Werkzeuge sowie Trainings- und Testdaten samt einer Anleitung zur Nutzung des Systems werden unter einer Open-Source-Lizenz veröffentlicht und stehen der Allgemeinheit in Kürze zur Verfügung.

Referenzen

Breuel, T. M., Ul-Hasan, A., Al-Azawi, M. A., & Shafait, F. (2013). High-performance OCR for printed English and Fraktur using LSTM networks. In *Document Analysis and Recognition (ICDAR), 2013,* 683-687.

Federbusch, M., Polzin, C., & Stäcker, T. (2013). *Volltext via OCR - Möglichkeiten und Grenzen: Testszenarien zu den Funeralschriften der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz. Erfahrungsbericht aus dem Projekt "Helmstedter Drucke Online" der Herzog August Bibliothek Wolfenbüttel/von Thomas Stäcker*. Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, Berlin.

Habermann, M. (2003). Der Sprachenwechsel und seine Folgen. Zur Wissensvermittlung in lateinischen und deutschen Kräuterbüchern des 16. Jahrhunderts. In: *Sprachwissenschaft 28*, 325–354.

Klein, W.-P. (2011). Die deutsche Sprache in der Gelehrsamkeit der frühen Neuzeit. Von der *lingua barbarica* zur *HaubtSprache*. In: Jaumann, Herbert (Hg.) *Diskurse der Gelehrtenkultur in der Frühen Neuzeit. Ein Handbuch*. de Gruyter, Berlin/New York, 465–516

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780.

Springmann, U., Najock, D., Morgenroth, H., Schmid, H., Gotscharek, A., & Fink, F. (2014). OCR of historical printings of Latin texts: problems, prospects, progress. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage,* 71-75.