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

Uwe Springmann[1], Anke Lüdeling[2], Felix Schremmer[2]

[1]Centrum für Informations- und Sprachverarbeitung, Ludwig-Maximilians-Universität München
[2]Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin

springmann@cis.uni-muenchen.de, anke.luedeling@rz.hu-berlin.de, felix.schremmer@gmx.de

1. Einführung

Wir stellen eine neue OCR-Methode vor, mit der es erstmals möglich ist, gedruckte Texte von der Inkunabelzeit (1450-1500) bis heute mit hoher Genauigkeit (größer 95% Zeichenerkennungsrate) in elektronischen Text zu verwandeln. Bisherige Versuche der Konversion von Inkunabeln lieferten keinen brauchbaren Text (Rydberg-Cox, 2009).

Die Methode beruht auf rekurrenten neuronalen Netzen mit langem Kurzzeitgedächtnis (Hochreiter und Schmidhuber, 1997), deren Anwendbarkeit auf OCR erstmals von Breuel u. a. (2013) beschrieben wurde. Durch den Vergleich von gedruckten Textzeilen mit einer diplomatischen Transkription („ground truth“), die das Netzwerk als Input erhält, werden die internen Parameter in einem automatischen Verfahren so eingestellt, dass nach einer hinreichenden Anzahl von Lernschritten eine Erkennung neuer Textzeilen mit hoher Genauigkeit möglich wird. Das Verfahren benötigt daher Trainingsdaten in Form einer diplomatischen Transkription, wie sie im diachronen RIDGES-Korpus zur Verfügung stehen. Ein trainiertes Modell kann dann auf Drucke mit gleicher oder ähnlicher Typographie angewendet werden. Die Resultate sind sprachunabhängig und setzen kein Lexikon voraus.

Damit eröffnet sich die Möglichkeit, das gedruckte Erbe auch bei sehr frühen Drucken maschinell in digitalen Text zu transformieren. Mögliche Anwendungen ergeben sich für die Suche (mit Trefferanzeige im Bild anstelle im OCR-Text) sowie, gegebenenfalls nach automatischer und manueller Nachkorrektur, für den automatisch unterstützten Aufbau von Korpora und Lexika.

2. Das RIDGES-Korpus

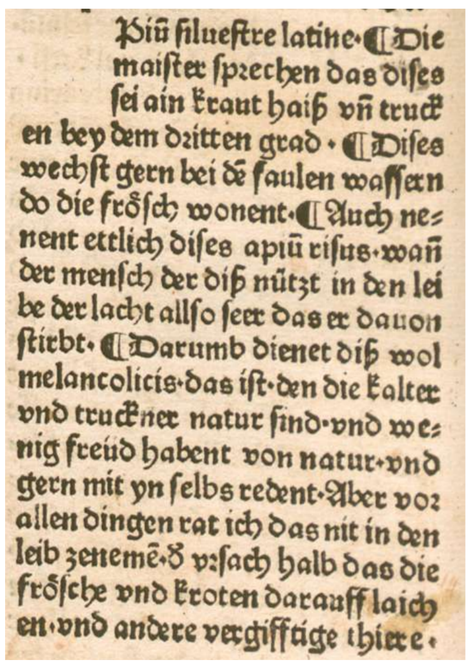
Als Trainingsdaten wurde die diplomatische Ebene des RIDGES-Korpus verwendet, das unter CC-BY verfügbar ist (Lüdeling u.a.). RIDGES steht für Register in Diachronic German Science; Ziel des RIDGES-Projekts ist die qualitative und quantitative Analyse der Entstehung eines deutschsprachigen wissenschaftlichen Registers. Es handelt sich dabei um derzeit 24 Kräutertexte, die zwischen 1487 und 1870 veröffentlicht wurden und von denen jeweils etwa 30 Seiten in mehreren Ebenen annotiert wurden. Damit steht zur Überprüfung der Anwendbarkeit der neuen OCR-Methode eine repräsentative Stichprobe zur Verfügung. Die OCR-Ergebnisse sind auch ihrerseits wieder für das Korpus nützlich, da erst über digital vorliegende Texte statistische Registeranalysen möglich werden.

3. Die Methode

Eine erfolgreiche OCR alter Drucke durchläuft die folgenden Schritte:

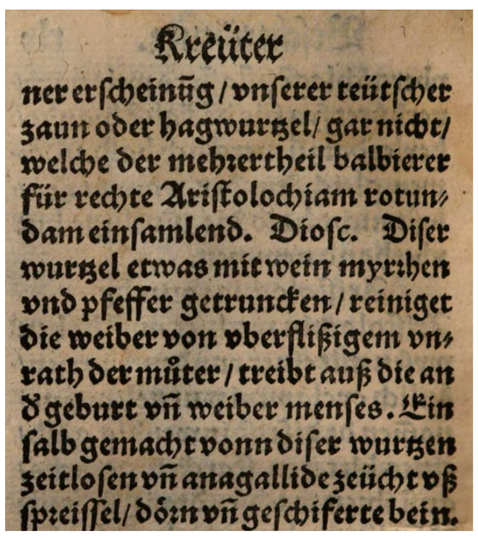
1. Beschaffung eines guten Scans (mind. 300 dpi in Graustufen oder Farbe)
2. Vorverarbeitung: Deskewing, Dewarping, Despeckling, Binarisierung, evtl. Zoning (Zerlegung des Seitenbildes in Subimages, z.B. bei Mehrspaltendruck)
3. Zerlegung der Textbereiche in Zeilenbilder
4. Diplomatische Erstellung einer zeilengerechten Transkription (Ground Truth)
5. Training eines Modells auf einer Trainingsmenge von Zeilenbildern mit zugehöriger Ground Truth
6. Anwendung des Modells zur Texterzeugung für das gesamte Druckwerk

Zwei Beispiele für die Anwendung von Modellen auf Testseiten geben die Abb. 1 und 2.



Piů ilueſtre latine · ¶ Die maiſter ſprechen das diſzs ſei ain kraut haiſz vñ truck en bey dem dritten grad · ¶ Diſes wechſt gern bei dẽ faulen waſſern do die frõſch wonent · ¶ Auch ne= nent ettlich diſes apnů riſus · wañ der menſch der diſz nützt in den lel be der lacht allſo ſeer das er dauon ſtirbt · ¶ Darumb dienet diſz wol melancolicis · das iſt · den die kalter vnd truckner natur ſind · vnd wt= nig freud habent von natur vnd gern mit yn ſelbs redent · Aber vor allen dingen rat ich das nit in den leib zenemẽð vrſach halb das die frõſche vnd kroten darauff laich en · vnd andere vergifftige nhiere ·

Abbildung 1: *Johann Wonnecke von Kaub (Johannes von Cuba): Gart der Gesundheit, Ulm 1487. Ergebnis der Anwendung eines auf 9 Seiten trainierten Modells (45.000 Lernschritte, Zeichenerkennungsrate 96,4%) auf eine vorher nicht gesehene Testseite.*



Kreüter ner erſcheinüg / vnſerer teütſcher zaun oder hagwurtzel / gar nicht / welche der mehrertheil balbierer für rechte Ariſtolochiam rotun= dam einſamlend. Dioſc. Diſer wurtzel etwas mit wein myrrhen vnd pfeffer getruncken / reiniget die weiber von vberfliſzigem vn= rath der můter / treibt auſz die an d geburt vñ weiber menſes. Ein ſalb gemacht vonn diſer wurtzen zeitloſen vñ anagallide zeücht vſz ſpreiſſel / dórn vñ geſchiferte bein.

Abbildung 2: *Adam von Bodenstein, Wie sich meniglich . . . , Basel 1557. Trainiert wurde auf 35 Seiten mit 32.000 Lernschritten, getestet auf 5 Seiten. Die Zeichenerkennungsrate liegt über 99%.*

4. Ergebnisse

Abb. 3 stellt die mit den trainierten Modellen auf den Testseiten der einzelnen Werke erreichten Zeichenerkennungsraten dar (angegeben ist der verbliebene Fehler, d.h. 1 - Zeichenerkennungsrate). Der mittlere Fehler liegt durchweg unter 5%. Es ist kein Trend zu schlechteren Ergebnissen bei älteren Drucken erkennbar, vielmehr hängt die Erkennungsgüte von der Qualität der Druckvorlage bzw. des Scans ab. Werke mit einer größeren Streuung der Ergebnisse haben auch einen schlechteren Mittelwert. Dies korreliert mit vergleichsweise schlechteren Scans, da zum Teil auf die bei Google Books vorhandenen, oft schon binarisierten Scans mit geringer Auflösung (150 dpi) zurückgegriffen wurde. Ein Nachtraining des Werks von 1557 ausgehend von besseren Scans (BSB anstatt Google Books) konnte beispielsweise den Fehler von 5% auf 1% reduzieren.

Über die Jahrhunderte hinweg ist es also möglich, bei guten Scans Zeichengenauigkeiten zwischen 95% und 99% bereits ohne Einsatz von Lexika und vor einer Nachkorrektur zu erhalten.

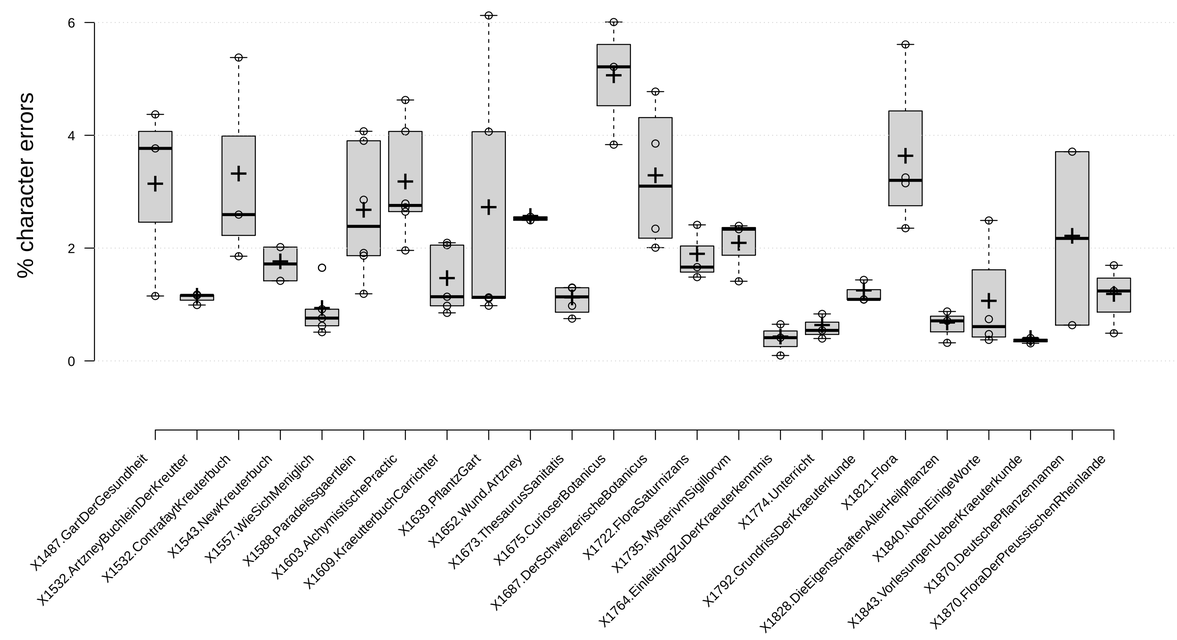


Abbildung 3: *Prozentuale Fehler der Zeichenerkennung pro Testseite. Die Ringe geben die Werte für einzelne Seiten an, das + Zeichen den Mittelwert, die Boxen die obere und untere Quartile und die Striche das Minimum und Maximum.*

5. Diskussion

Der Knackpunkt der Verwendbarkeit der neuen Methode liegt im Aufwand, der für die manuelle Transkription einer diplomatischen Ground Truth als Voraussetzung für das Training erbracht werden muss. Die großflächige und möglichst automatisierte Anwendung setzt also Modelle voraus, die einen weiten Bereich an Typographien abdecken. Für den Bereich moderner Druckschriften, die weitgehend auch als Computerfonts verfügbar sind, kann das Training ausgehend von elektronischem Text effizient durchgeführt werden, indem man zum vorhandenen Text mit einer Vielzahl von Schrifttypen verrauschte Bilder erzeugt, die dann direkt für das Training eingesetzt werden können. Da solche Fonts für historische Schriften bisher nicht in gewünschter Qualität und Anzahl zur Verfügung stehen, bleibt vorerst nur der Weg der manuellen Transkription, wie von Springmann u. a. (2014) gezeigt wurde. Mit zunehmendem Vorrat an Ground Truth besteht die Hoffnung, dass auf gemischten Schriftarten trainierte Modelle eine zunehmende Anwendbarkeit gewinnen.

Für einzelne Werke oder Buchreihen, die mit einheitlicher Schrifttype gedruckt wurden, ist das Verfahren jetzt schon ein Durchbruch in der Herstellung von elektronischem Text. Insbesondere wird bereits durch das unkorrigierte OCR-Resultat die Möglichkeit einer fehlertoleranten Suche mit Anzeige der Fundstellen im Bild geschaffen. Für den Aufbau von fehlerkorrigierten Korpora spart das initiale Training auf ein paar Seiten bei der anschließenden automatisch erfolgenden Konversion des Textes viel Zeit, da nur noch nachkorrigiert werden muss.

Die Effizienz der Nachkorrektur hängt dabei auch von verwendeten Werkzeugen ab, die gezielt zu vermuteten Fehlerserien hinführen und eine Stapelkorrektur ganzer Serien ermöglichen. Ein solches Werkzeug unter dem Namen PoCoTo wurde am CIS als Prototyp entwickelt und steht unter einer Open-Source-Lizenz zur Verfügung (Vobl u. a., 2014).

Unter dem Projektnamen „ocrocis“ (Springmann und Kaumanns) steht ein am CIS entwickelter Wrapper für Breuel’s ocropy-System bereit, mit dem das OCR-System unter Linux und MacOS installiert und ausprobiert werden kann.

Literatur

[Breuel u. a. 2013] Breuel, Thomas M. ; Ul-Hasan, Adnan ; Al-Azawi, Mayce A. ; Shafait, Faisal: High-Performance OCR for Printed English and Fraktur using LSTM Networks. In: *2th International Conference on Document Analysis and Recognition (ICDAR), 2013* IEEE (Veranst.), 2013, S. 683–687

[Hochreiter und Schmidhuber 1997] Hochreiter, Sepp ; Schmidhuber, Jürgen: Long short-term memory. In: *Neural computation* 9 (1997), Nr. 8, S. 1735–1780

[Lüdeling u.a.] : *RIDGES – Register in Diachronic German Science*. http://korpling.german.hu-berlin.de/ridges/index_de.html

[Rydberg-Cox 2009] Rydberg-Cox, Jeffrey A.: Digitizing Latin incunabula: Challenges, methods, and possibilities. In: *Digital Humanities Quarterly* 3 (2009), Nr. 1

[Springmann und Kaumanns] Springmann, Uwe ; Kaumanns, David: *ocrocis – a high accuracy OCR method to convert early printings into digital text*. https://code.google.com/p/cistern/

[Springmann u. a. 2014] Springmann, Uwe ; Najock, Dietmar ; Morgenroth, Hermann ; Schmid, Helmut ; Gotscharek, Annette ; Fink, Florian: OCR of historical printings of Latin texts: problems, prospects, progress. In: *DATeCH*, 2014, S. 71–75

[Vobl u. a. 2014] Vobl, Thorsten ; Gotscharek, Annette ; Reffle, Uli ; Ringlstetter, Christoph ; Schulz, Klaus U.: PoCoTo - an Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts. In: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, 2014 (DATeCH ’14), S. 57–61