OCR für alte Drucke

Uwe Springmann

Einleitung

Unter allen kulturellen Artefakten nimmt die Schrift nach Menge und Bedeutung zweifellos den ersten Rang ein. Das schriftlich fixierte kulturelle Erbe ist in Manuskripten, der Menge nach aber vor allem in gedruckten Büchern gespeichert, die seit der Medienrevolution durch die Erfindung des Drucks mit beweglichen Metalllettern (1453: Gutenbergs Bibeldruck) vorliegen. Nach mehr als einem halben Jahrtausend steht nun die nächste Revolution an, die unser vorhandenes Wissen in ein neues Medium transformieren, gleichzeitig aber auch neuen Verwendungen zugänglich machen wird: Eine Reihe von Digitalisierungsprojekten, allen voran Google Books, ist seit Jahren damit beschäftigt, digitale Seitenbilder dieser Bücher durch Anwendung der Buchscantechnologie zu erzeugen. Es ist absehbar, dass der Großteil aller jemals hergestellten Buchtitel in einigen Jahren vollständig digitalisiert und für jedermann zugreifbar im Internet vorliegen wird.

Damit scheint sich der Traum von der universalen Bibliothek, die Vision des freien Zugriffs auf das gesamte Menschheitswissen, demnächst zu verwirklichen. Es gibt jedoch eine Einschränkung: Bildseiten können nicht maschinell durchsucht werden und erfordern zu ihrer Interpretation ein Augenpaar mit dahinter geschaltetem Gehirn. Die echte universale Bibliothek liegt erst dann vor, wenn ihre Texte auch in maschinenverarbeitbarer Form zur Verfügung stehen, damit Fundstellen, Querverweise, Abhängigkeiten und die ganze Breite des intellektuellen Diskurses auf Knopfdruck zur Verfügung stehen.

Die Anfang des 20. Jahrhunderts erfundene Technologie der optischen Zeichenerkennung (Optical Character Recognition, OCR) ist in diesem Zusammenhang von zentraler Bedeutung. Der folgende Artikel gibt Auskunft über den in den vergangenen Jahren erzielten Durchbruch auf dem Gebiet der OCR alter Drucke auf der Basis von rekurrenten neuronalen Netzen.

Schwierigkeiten bei der OCR alter Drucke

Auf modernen Drucken funktioniert die OCR mit proprietären (Marktführer: ABBYY[1]) oder Open-Source Produkten (Tesseract[2]) mit Zeichenerkennungsraten von über 99 % sehr gut. Diese Drucke zeichnen sich durch ein hohes Signal-zu-Rausch-Verhältnis aus, bei dem die Schriftzeichen klar erkennbar sind und sich gut von einem einheitlich hellen Hintergrund ohne Störungen wie etwa Verschmutzungen, Anstreichungen, Bräunungen, durchscheinende Rückseiten etc. abheben. Zudem sind die OCR-Engines auf einer Vielzahl von Schriften trainiert, die im modernen Druck üblich und als Computerfonts leicht verfügbar sind. Alle diese Voraussetzungen gelten jedoch für alte Drucke nicht,

[1] http://www.frakturschrift.com/de:start

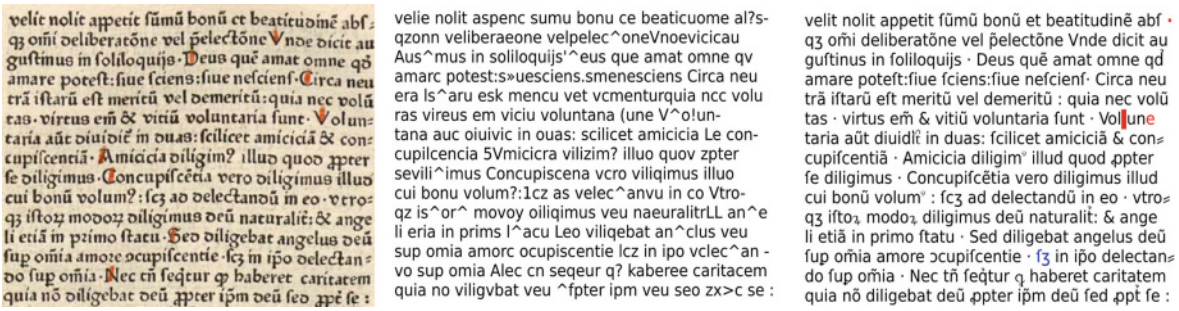
[2] https://github.com/tesseract-ocr/tesseract/wiki

DOI 10.1007/s00287-016-1004-3
© Springer-Verlag Berlin Heidelberg 2016

Uwe Springmann
Ludwig-Maximilians-Universität,
Centrum für Informations- und Sprachverarbeitung (CIS),
Oettingenstraße 67, 80538 München
E-Mail: springmann@cis.uni-muenchen.de

*Vorschläge an Prof. Dr. Frank Puppe
<puppe@informatik.uni-wuerzburg.de>
oder an Dr. Brigitte Bartsch-Spörl
<brigitte@bsr-consulting.de>

Alle „Aktuellen Schlagwörter" seit 1988 finden Sie unter:
http://www.is.informatik.uni-wuerzburg.de/as



velie nolit aspenc sumu bonu ce beaticuome al?s-
qzonn veliberaeone velpelec^oneVnoevicicau
Aus^mus in soliloquijs'^eus que amat omne qv
amarc potest:s»uesciens.smenesciens Circa neu
era Is^aru esk mencu vet vcmenturquia ncc volu
ras vireus em viciu voluntana (une V^o!un-
tana auc oiuivic in ouas: scilicet amicicia Le con-
cupilcencia 5Vmicicra vilizim? illuo quov zpter
sevili^imus Concupiscena vcro viliqimus illuo
cui bonu volum?:1cz as velec^anvu in co Vtro-
qz is^or^ movoy oiliqimus veu naeuralitrLL an^e
li eria in prims l^acu Leo viliqebat an^clus veu
sup omia amorc ocupiscentie lcz in ipo vclec^an -
vo sup omia Alec cn seqeur q? kaberee caritacem
quia no viligvbat veu ^fpter ipm veu seo zx>c se :

velit nolit appetit ſũmũ bonũ et beatitudinẽ abſ ·
qȝ oñi deliberatõne vel ꝑelectõne Vnde dicit au
guſtinus in ſoliloquijs · Deus quẽ amat omne qd
amare poteſt:ſiue ſciens:ſiue neſcienſ· Circa neu
trã iſtarũ eſt meritũ vel demeritũ : quia nec volũ
tas · virtus eñ & vitiũ voluntaria ſunt · Vol une
taria aũt diuidiť in duas: ſcilicet amiciciã & con-
cupiſcentiã · Amicicia diligim⁹ illud quod ꝑpter
ſe diligimus · Concupiſcẽtia vero diligimus illud
cui bonũ volum⁹ : ſcȝ ad delectandũ in eo · vtro-
qȝ iſto⁊ modo⁊ diligimus deũ naturaliť: & ange
li etiã in primo ſtatu · Sed diligebat angelus deũ
ſup oñia amore ꝺcupiſcentie · ſcȝ in ipõ delectan-
do ſup oñia · Nec tñ ſeq̃tur q̃ haberet caritatem
quia nõ diligebat deũ ꝑpter ipm deũ ſed ꝑpť ſe :

velit nolit appetit ſũmũ bonũ et beatitudinẽ abſ ·
qȝ oñi deliberatõne vel ꝑelectõne Vnde dicit au
guſtinus in ſoliloquijs · Deus quẽ amat omne qd
amare poteſt:ſiue ſciens:ſiue neſcienſ· Circa neu
trã iſtarũ eſt meritũ vel demeritũ : quia nec volũ
tas · virtus eñ & vitiũ voluntaria ſunt · Volune
taria aũt diuidiť in duas: ſcilicet amiciciã & con-
cupiſcentiã · Amicicia diligim⁹ illud quod ꝑpter
ſe diligimus · Concupiſcẽtia vero diligimus illud
cui bonũ volum⁹ : ſcȝ ad delectandũ in eo · vtro-
qȝ iſto⁊ modo⁊ diligimus deũ naturaliť: & ange
li etiã in primo ſtatu · Sed diligebat angelus deũ
ſup oñia amore ꝺcupiſcentie · ſcȝ in ipõ delectan-
do ſup oñia · Nec tñ ſeq̃tur q̃ haberet caritatem
quia nõ diligebat deũ ꝑpter ipm deũ ſed ꝑpť ſe :

Abb. 1 Ein Ausschnitt einer Seite aus Vincent von Beauvais, Speculum Naturale, Straßburg, 1476* (links) *zusammen mit dem OCR-Ergebnis von ABBYY Finereader 11* (Mitte) *und einem trainierten OCRopus-Modell* (rechts). *Das Finereader-Ergebnis hat eine Zeichenerkennungsrate von 68 %, bei OCRopus sind es 98 % (die verbleibenden Fehler sind:* 1. Zeile: *abſ = → abſ ·*, 6. Zeile: *Volun= → Vol une*, 3. Zeile *von unten: ſcȝ → ſȝ)

wobei die Verwendung von druckerspezifischen Schriftarten ein Haupthindernis für gute Erkennung darstellt. Abbildung 1 zeigt ein Beispiel einer Inkunabel (Druck des 15. Jahrhunderts) und rechts daneben das OCR-Ergebnis von ABBYY Finereader (Zeichenerkennungsrate 68 %), bei dem fast kein Wort richtig erkannt wurde. Ein Training auf den verwendeten Schriftarten ist daher unumgänglich, was z. B. mit Tesseract möglich, aber sehr aufwendig ist: Alle verwendeten Glyphen eines Werks müssen händisch aus den Bildseiten ausgeschnitten werden, um damit die verwendete Schriftart zu rekonstruieren. Anschließend wird ein vorhandener Text mithilfe dieses künstlichen Fonts wieder in ein verrauschtes Bild umgesetzt und Tesseract mit diesen Daten trainiert.[3] Dadurch ist es möglich, Erkennungsraten bei Inkunabeln von über 90 % zu erzielen [5], was jedoch für viele Anwendungen immer noch nicht ausreicht und einen erheblich Nachkorrekturbedarf nach sich zieht.

Neue OCR-Methoden auf Basis rekurrenter neuronaler Netze mit LSTM-Architektur

Zum Glück stehen mittlerweile auch effizientere Methoden für die OCR zur Verfügung, die sowohl einfacher zu trainieren sind als auch erheblich bessere Erkennungsqualität liefern. Diese Methoden beruhen auf der LSTM(Long Short Term Memory)-Architektur für rekurrente neuronale Netze, die auf Hochreiter und Schmidhuber [4] zurückgeht und mit großem Erfolg in der Mustererkennung eingesetzt wurde, u. a. auch bei der Erkennung von Handschriften und mittelalterlichen Manuskripten [2, 3].

Während Tesseract naheliegenderweise einzelne Zeichen als atomare Textbausteine betrachtet und prototypische Zeichen durch bestimmte Charakteristiken beschreibt, die es im gedruckten Text wiederzuentdecken gilt, beruht die Erkennung mittels LSTM-Netzen auf einem radikalen Neuansatz: Hier werden ganze Textzeilen in jeweils einen Pixel breite vertikale Streifen zerlegt, was pro Zeichen zu einer Übersegmentierung mit bis zu 30 Streifen führt. Die Klassifikation von wiederkehrenden Streifenfolgen zu einzelnen Zeichen geschieht dann automatisch während des Trainingsprozesses, bei dem jeweils eine gedruckte Textzeile zusammen mit ihrer zugehörigen Transkription (auch *ground truth* genannt) dem Netz als Input gegeben wird. Die Abweichung zwischen vorhergesagten Streifenfolgen (die am Ende einem Zeichen entsprechen) und dem wirklichen Zeichen wird an das Netz zurückgegeben, das daraufhin die Gewichte für die Verbindungen zwischen seinen Netzknoten so adaptiert, dass diese Abweichung in Zukunft minimiert wird. Da jedoch jede Trainingszeile wieder anders aussieht, besteht die Gefahr, dass die gelernten Zusammenhänge bis zum nächsten Auftreten ähnlicher Streifenmuster überschrieben und damit vergessen werden. Hier kommt LSTM ins Spiel: Jeder Netzwerkknoten ist von einer Struktur (sog. *Gates*) umgeben, die selektiv das Aufnehmen neuer Informationen zulässt bzw. verhindert. Damit sind stets nur einige Knoten aktiv am Lernvorgang beteiligt, während andere ihre Information unverändert behalten und das bisher Gelernte im Gedächtnis bewahren. Ähnliche Mechanismen gibt es für Vergessen und Ausgabe.

[3] Das Vorgehen wird hier beschrieben: http://emop.tamu.edu/outcomes/Franken-Plus

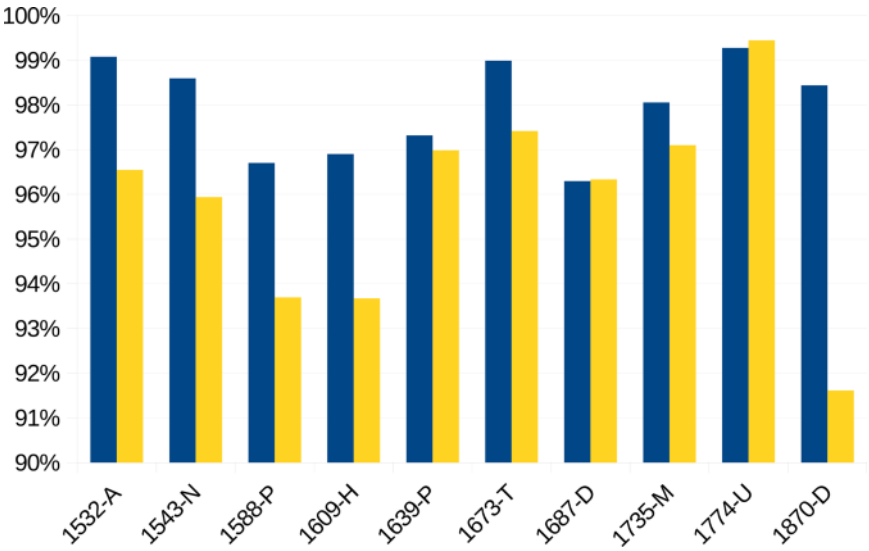


Abb. 2 Erkennungsraten mit individuellen* (dunkel) *und einem gemischten* (hell) *Modell für zehn deutsche Bücher in Fraktur. Die Zahlen in den Spaltenunterschriften geben das Erscheinungsjahr an. Dasselbe (gemischte) Modell, das auf zehn anderen Büchern trainiert wurde, erzeugt gute Erkennungsraten über einen Druckzeitraum von 338 Jahren

Die Gates kann man also als Leitungen verstehen, die den Inhalt einer Gedächtniszelle beeinflussen und durch Ventile selektiv geöffnet und geschlossen werden. Auch dies geschieht während des Lernvorgangs automatisch. Am Ende wird eine Zeichenfolge ausgegeben, die eine Vorhersage des Netzes über den Inhalt einer gedruckten Zeile darstellt. Tatsächlich wird sogar eine Konfidenzmatrix berechnet, die jeder Position einer Zeile eine Wahrscheinlichkeitsverteilung über alle möglichen Zeichen (den *Setzkasten*) zuordnet, wobei die tatsächlich ausgegebenen Zeichen die jeweils höchste Konfidenz von allen Zeichen an einer Position haben.

Breuel et al. [1] haben diese Algorithmen in ihre Open-Source OCR-Software *OCRopus* aufgenommen, die auf Github zur Verfügung steht,[4] und gezeigt, dass mit einem trainierten Modell sehr gute Erkennungsraten auf englischen Drucken und deutschen Frakturdrucken des 19. Jahrhunderts erzielt werden können.

Training neuronaler Netze

Auch bei OCRopus reichen die auf modernen Schriftarten trainierten Standardmodelle zur Erkennung nicht aus, um alte Drucke gut zu erkennen. Ein Ausschneiden von Glyphen zur Rekonstruktion historischer Typen ist zwar nicht notwendig, aber für das Training auf realen Drucken wird eine gewisse Anzahl an transkribierten Zeilen benötigt. Jedoch erhält man bereits mit 100 bis 200 transkribierten Zeilen sehr gute Ergebnisse von meist über 95 % Zeichenerkennungsrate [6], abhängig von der Güte der Bildvorlage und der Segmentierung der Bildseite in Text- und Nichttext-Bereiche. Der Zeilenvorrat einer Trainingsmenge wird dabei zufällig ausgewählt und nach einer bestimmten Anzahl von Lernschritten (jeder Schritt entspricht der Präsentation einer Zeile mit ihrer Transkription) wird ein Modell herausgeschrieben. Nach einigen tausend Schritten liegt eine Reihe von Modellen vor, aus denen dann das beste Modell durch Evaluation auf einer Testmenge herausgesucht wird. Die Inkunabel aus Abb. 1 konnte mit einem auf diese Weise trainierten OCRopus-Modell mit 98 % Zeichenerkennungsrate in elektronischen Text verwandelt werden.[5]

Mit der Möglichkeit des Trainings individueller Schriftarten steht nunmehr eine Methode bereit, die eine OCR mit bisher unerreichten Erkennungsraten selbst der frühesten Bücher der neuzeitlichen Druckgeschichte ermöglicht. Jedoch wäre es immer noch ein gewaltiger Aufwand, wenn man jede einzelne Drucktype trainieren müsste. Wenn man auch den maschinellen Aufwand beim Training vernachlässigen kann, so bleibt immer noch die Notwendigkeit, eine gewisse Textmenge manuell zu transkribieren. Leider hat sich gezeigt, dass

[4] https://github.com/tmbdev/ocropy

[5] Eine Anleitung für das Training eigener Modelle mit OCRopus ist hier verfügbar: http://cistern.cis.lmu.de/ocrocis/ und ein ausführlicher Workshop zur OCR historischer Drucke steht zusammen mit Beispieldateien zum Ausprobieren auf Github bereit: https://github.com/cisocrgroup/OCR-Workshop.

vorhandene OCR-Modelle selbst bei oberflächlich sehr ähnlich aussehenden Schriften oftmals viel schlechtere Ergebnisse als individuell trainierte Modelle liefern. Eine Möglichkeit, diese Typografiebarriere zu durchbrechen, stellt das Training von gemischten Modellen dar, die auf der kombinierten Trainingsmenge einer Mehrzahl von Drucken mit unterschiedlichen Schriftarten trainiert werden. Abbildung 2 zeigt das Ergebnis der OCR auf zehn deutschen Büchern, die zwischen 1532 und 1870 in Fraktur gedruckt wurden. Gezeigt werden Erkennungsraten mit individuellen (dunkel) und einem gemischten (hell) Modell, das auf zehn anderen Büchern aus demselben Zeitraum trainiert wurde.[6] Die Ergebnisse liegen selbst beim gemischten Modell zum Teil erheblich über 90 %, und übertreffen im Fall von 1774-U sogar das individuelle Modell (offensichtlich enthielt die Trainingsmenge der gemischten Modelle etliches Material, das mit derselben Type gedruckt wurde). Wenn diese Ergebnisse im Einzelfall noch nicht ausreichen sollten, kann man sich ausgehend vom erkannten Text durch manuelle Korrekturen leicht individuelles Trainingsmaterial erstellen und nachtrainieren.

Wie man die Landschaft der historischen Drucktypen sinnvoll nach Perioden, Regionen und Druckerwerkstätten (*Offizinen*) einteilt, um möglichst allgemein verwendbare gemischte Modelle zu trainieren, ist derzeit noch eine offene Frage. Das obige Ergebnis deutet zumindest darauf hin, dass ein solches Vorgehen sinnvoll ist.

[6] Siehe http://arxiv.org/abs/1608.02153

Offene Probleme

Aber auch jenseits der eigentlichen Erkennung sind noch offene Probleme vorhanden. Die automatische Segmentierung von Bildseiten funktioniert bisher nicht befriedigend, und alle Fehler in diesem frühen Verarbeitungsschritt wirken sich später negativ auf die Erkennungsrate aus. Es bleibt abzuwarten, ob auch für diesen Bereich neuronale Netze einen Beitrag liefern werden.

Abschließend ist noch die Weiterverarbeitung des erkannten Textes zu nennen, und zwar sowohl die Fehlerkorrektur als auch die Standardisierung bzw. Normalisierung, sodass man auch auf Texten in historischer Schreibweise modern buchstabierte Suchanfragen durchführen kann. Neben regelbasierten und statistischen Ansätzen könnte auch hier das Training neuronaler Netze, diesmal mit Trainingspaaren von Zeilen in fehlerbehafteter und richtiger Schreibweise, bzw. historischer und moderner Schreibweise, einen Fortschritt bringen. Erste Ansätze dazu gibt es bereits.

Literatur

1. Breuel TM, Ul-Hasan A, Al-Azawi MA, and Shafait F (2013) High-Performance OCR for Printed English and Fraktur Using LSTM Networks. In: 2th International Conference on Document Analysis and Recognition (ICDAR). IEEE, pp 683–687
2. Fischer A, Wuthrich M, Liwicki M, Frinken V, Bunke H, Viehhauser G, Stolz M (2009) Automatic Transcription of Handwritten Medieval Documents. In: 15th International Conference on Virtual Systems and Multimedia (VSMM'09). IEEE, pp 137–142
3. Graves A, Liwicki M, Fernández S, Bertolami R, Bunke H, Schmidhuber J (2009) A novel connectionist system for unconstrained handwriting recognition. IEEE T Pattern Anal 31(5):855–868
4. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
5. Kirchner F, Dittrich M, Beckenbauer P, Nöth M (2016) OCR bei Inkunabeln – Offizinspezifischer Ansatz der UB Würzburg. ABI Tech 36(3):178–188
6. Springmann U, Fink F, Schulz KU (2016) Automatic quality evaluation and (semi-) automatic improvement of mixed models for OCR on historical documents. http://arxiv.org/abs/1606.05157. Letzter Zugriff: 26.10.2016