Über den Mehrwert der Vernetzung von OCR-Verfahren zur Erfassung von Texten des 17. Jahrhunderts

,
boenig@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften - Berlin, Deutschland

,
wuerzner@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften - Berlin, Deutschland

,
binder@informatik.hu-berlin.de
Berlin-Brandenburgische Akademie der Wissenschaften - Berlin, Deutschland

,
springmann@cis.uni-muenchen.de
Centrum für Informations- und Sprachverarbeitung - Ludwig-Maximilians-Universität München, Deutschland

Einleitung

Dieser Beitrag stellt eine neuartige Methode zur optischen Zeichenerkennung (*Optical Character Recognition*, OCR) speziell für Textvorlagen des 17. Jahrhunderts vor. Anstatt ein neues OCR-Verfahren zu entwickeln, werden zwei etablierte Open-Source-Lösungen genutzt. Die Ausgaben der Programme werden computergestützt kombiniert, um so eine möglichst genaues Textergebnis zu erhalten. Die Besonderheiten und die Güte der Methode wird anhand der Texterfassung von Gelegenheitsgedichten von Simon Dach illustriert.

OCR

OCR bezeichnet die Gesamtheit von Verfahren, die in der Lage sind, aus Rastergrafiken Schriftzeichen zu erkennen. Der Begriff wird sowohl für die eigentliche Mustererkennung als auch für den gesamten Prozess der Bildverarbeitung verwendet. Letzterer gliedert sich normalerweise in drei Schritte: **1. Bildoptimierung**: Diese besteht aus der Bitonalisierung der Digitalisate, ihrer Begradigung (sog. *Deskewing*) und aus der Entfernung von Artefakten (sog. *Despeckling*). Außerdem können beim Scannen entstandene Wellen in einzelnen Zeilen automatisch begradigt werden (sog. *Dewarping*). **2. Strukturerkennung** (*Optical Layout Recognition*, OLR): Die einzelnen Seiten werden u. a. in Spalten, Absätze und Zeilen gegliedert. **3. Mustererkennung** (OCR): Für diese Aufgabe gibt es verschiedene Lösungsvorschläge sowohl im kommerziellen wie auch im Open-Source-Bereich. Besonders verbreitet sind die Software *FineReader* der Firma ABBYY sowie *BITAlpha* aus dem Hause Tomasi, die u. a. von Bibliotheken eingesetzt werden. Die bekanntesten Open-Source-Lösungen sind das ursprünglich von Hewlett-Packard entwickelte und heute von Google betreute *Tesseract* (GitHub 2016a) und das ursprünglich am DFKI Kaiserslautern entwickelte *OCRopus* (GitHub 2016b).

Grundsätzlich lassen sich bei OCR zwei unterschiedliche Erkennungsansätze unterscheiden: zeichenorientierte Verfahren wie Tesseract vergleichen das Bild eines Zeichens Pixel für Pixel mit einer Datenbasis (dem sog. Modell) und geben das ähnlichste Zeichen zurück. Sequenzorientierte (segmentierungsfreie) Verfahren wie OCRopus legen ein Raster fester Größe über eine Zeile und bestimmen anhand der Folgen der einzelnen Spalten, repräsentiert als Bitvektoren (0 entspricht weiß, 1 schwarz) die wahrscheinlichste Zeichensequenz.

Gelegenheitsgedichte

Unsere Studie beschäftigt sich mit OCR am Beispiel von Gelegenheitsgedichten des 17. Jahrhunderts, denen durch die von Segebrecht (1977) initiierte literaturwissenschaftliche Neubewertung eine zunehmende kulturgeschichtliche Bedeutung zukommt (vgl. Klöker 2010: 39). Der Zugriff auf diese Drucke wurde durch das VD17 (HAB 2007-2016) und durch das *Handbuch des personalen Gelegenheitsschrifttums in europäischen Bibliotheken und Archiven* (Garber 2001-2013) erleichtert. Dennoch kann ein digitales Korpus für diese Textsorte heute nur als Desiderat wahrgenommen werden. Für Werke von Simon Dach ist die Ausgangslage scheinbar besser: Mit der digitalisierten vierbändigen Ausgabe von Ziesemer (Ziesemer 1936-1938) steht ein großer Teil der heute bekannten Gedichte zur Verfügung (vgl. auch Dach o. J.; TextGrid 2015). Jedoch trübt sich dieser Eindruck beim textkritischen Blick.

111 Funeralschriften Simon Dachs wurden im Verlauf des DFG-Pilotprojektes zum *OCR-Einsatz bei der Digitalisierung der Funeralschriften der Staatsbibliothek zu Berlin* (2009-2011) (Federbusch / Polzin 2013) digitalisiert und per OCR erfasst. Die in der vorliegenden Studie genutzten Drucke zeichnen sich dahingehend aus, dass eine einheitliche Schrifttype sowie ein einfaches Layout vorliegen. Im Unterschied zu Texten des 18. und 19. Jahrhunderts war für diese Drucke noch ein relativ hoher manueller Aufwand erforderlich. Die Schrifttypen weisen daher eine vergleichsweise hohe Varianz bzgl. ihrer Form auf. Die 111 Trauergedichte weisen eine Textgenauigkeit von bis zu 95% auf. Der Schwerpunkt

der folgenden Studie liegt auf der Entwicklung und Prüfung von Methoden, die perspektivisch eine korrektere Übertragung der Textquellen aus dem 17. Jahrhundert liefern soll.

Arbeitsablauf

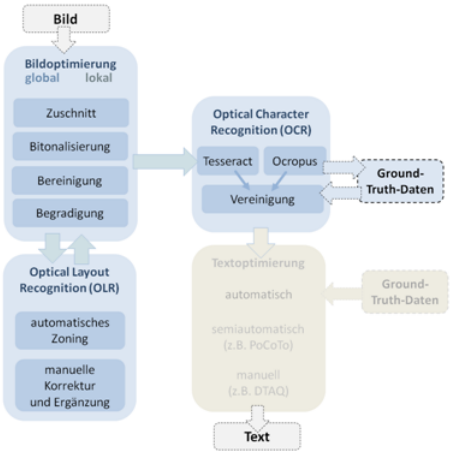

***Abb. 1**: Modell eines vollständigen Erfassungsworkflows (diese Studie betrifft die eingefärbten Stationen).*

Abbildung 1 gibt einen Überblick über den Arbeitsablauf der hier vorgestellten Methode. Im Unterschied zu existierenden Workflows unterteilt unser Vorschlag die Bildoptimierung in zwei Phasen: 1. *global*: Das komplette Digitalisat wird beschnitten, binarisiert, begradigt und von Artefakten befreit. Danach findet die Optische Layouterkennung (OLR) statt. 2. *lokal*: Die identifizierten Textzonen werden aus dem Bild der Seite ausgeschnitten und nochmals begradigt. Dadurch wird die häufig zu beobachtende Trapezform der Digitalisate, die durch Scannen von Büchern ohne Auftrennen des Buchrückens entsteht, behandelt. Die Bilder für die einzelnen Zonen werden anschließend in Zeilen zerschnitten und den OCR-Engines übergeben.

Unser Vorgehen bei der OCR orientiert sich an der manuellen Texterfassung per *Double Keying*: Dabei werden Texte von zwei unabhängigen Erfassern transkribiert. Im Vergleich der beiden Textversionen werden die Unterschiede ermittelt und die korrekte Version ausgewählt. Um den Genauigkeitsgewinn durch die Mehrfacherfassung zu erhöhen, wurden zwei paradigmatisch verschiedene OCR-Verfahren, Tesseract und OCRopus, mit unterschiedlichen Stärken und Schwächen eingesetzt. Beide Open-Source-Programme erlauben ein Training auf die vorwendeten Typen und die Anwendung spezifischer OCR-Modelle. Dies ist wie Springmann et al. (2015) zeigen ein wesentlicher Vorteil gegenüber den meisten Closed-Source-Lösungen, da die mitgelieferten OCR-Modelle insbesondere für frühe Druckerzeugnisse bzw. gebrochene Schriften sehr schlechte Ergebnisse bzgl. der Textgenauigkeit liefern. Die automatische Vereinigung der beiden Textversionen findet im Wesentlichen auf Basis einer Textdifferenzberechnung mit Hilfe von *diff* (Hunt / McIlroy 1976) statt, wobei im Falle von Unterschieden verschiedene Bewertungsheuristiken zur Bestimmung der *korrekten* Textversion eingesetzt werden. Das skizzierte Vorgehen erlaubt auch die Kombination von mehr als zwei Textversionen sowie den anschließenden Einsatz von OCR-Nachkorrekturverfahren (vgl. z. B. Vobl et al. 2014).

Evaluation

Die Güte der hier vorgestellten Methode wird anhand der Volltexterfassung von Funeralschriften Simon Dachs (vgl. 1.2) evaluiert. Dabei konzentriert sich die Evaluation auf drei Punkte:

- Welchen Einfluss hat die Wahl der Binarisierungsmethode auf die Textgenauigkeit?
- Wie groß ist der Unterschied zwischen einem Standardmodell und einem speziell für die zu erfassenden Texte trainierten Modell bzgl. der Textgenauigkeit?
- Kann die Vereinigung zweier durch OCR erzeugter Texte die Textgenauigkeit erhöhen?

Ein typisches Beispiel für die Untersuchungsgrundlage sowie die entsprechenden OCR-Ausgaben gibt Abbildung 2.

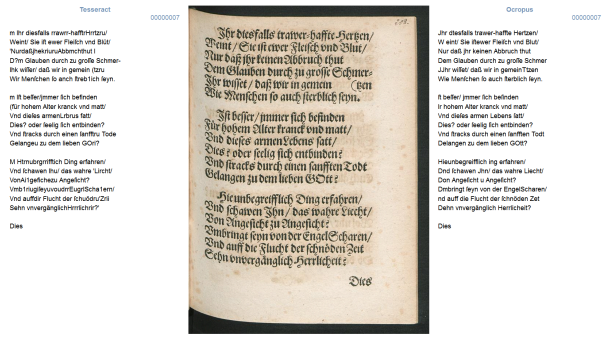

***Abb. 2**: Vergleich der OCR-Ergebnisse.*

Material

Ground Truth

Voraussetzung für die Evaluation und das Modelltraining ist fehlerfreier Volltext (*Ground Truth*). Um für die Studie entsprechende Daten zu gewinnen, wurde eine manuelle Korrektur aller 111 Texte vorgenommen. Die Korrektur schloss nicht nur die Text-, sondern auch die datenstrukturelle Ebene ein. Der Aufwand belief sich auf 150 Stunden. Im Ergebnis liegen alle

Texte im DTA-Basisformat vor und sind über die Qualitätssicherungsplattform DTAQ zugänglich.

Materialauswahl

Für das Training der spezifischen OCR-Modelle wurden 30 Seiten Ground-Truth zufällig ausgewählt. Für die Evaluation der Modelle wurden 25 andere zufällig ausgewählte Seiten verwendet.

Referenzlexikon

Zur Vereinigung beider OCR-Versionen wurde ein Referenzlexikon gültiger historischer Schreibungen des 17. Jahrhunderts herangezogen. Dazu wurden Wortformen (*n*=217067) aus DTA-Texten dieses Zeitraums extrahiert.

Durchführung

Vorverarbeitung

Für Beschneidung und Begradigung wurde das Programm *Scantailor* (GitHub 2016 a) eingesetzt. Für die Binarisierung, Artefaktbereinigung und Zeilenglättung wurde sowohl Scantailor als auch das in OCRopus enthaltene Werkzeug *nlbin* verwendet.

OLR

Die einzelnen Textzonen (Abschnitte und Kustoden) wurden mit Hilfe von *Leptonica* (Bloomberg 2001-2015) lokalisiert und manuell nachkorrigiert. Für die Untergliederung der Zonen in Zeilen wurde ebenfalls Leptonica eingesetzt.

OCR

Die Zeichenerkennung erfolgte sowohl mit OCRopus als auch mit Tesseract. Die erste Versuchsreihe basierte auf mitgelieferten Modellen. Für die zweite Versuchsreihe wurden die OCR-Programme mit Ground-Truth-Daten trainiert. Für das Training der OCRopus-Modelle wurde OCRopus eingesetzt. Dabei wurde für das Training aus Gründen der Modellvergleichbarkeit eine feste Anzahl von Iterationsschritten (*n*=30000) festgelegt. Die Tesseract-Modelle wurden mit Hilfe von *VietOCR* erstellt.

Textvereinigung

Die Textvereinigung wurde in *Python* mit Hilfe des Moduls *difflib* implementiert. Neben dem Referenzlexikon standen zur Konfliktauflösung auch die von den OCR-Programmen zurückgelieferten Konfidenzen auf Zeichenebene zur Verfügung. Waren sich die beiden Engines bzgl. eines Wortes bzw. einer Textsequenz uneins, wurde zunächst dem Wort Vorrang gegeben, dass sich im Referenzlexikon befindet. Konnte dort keine der beiden Versionen gefunden werden, wurde die Entscheidung auf Basis der Konfidenzwerte getroffen.

Qualitätsmessung

Die Bestimmung der Textqualität erfolgte durch Messung des Anteils falsch erkannter Zeichen (Fehlerrate in Prozent) im Vergleich zum fehlerfreien Volltext.

Ergebnisse und Diskussion

Tabelle 1 gibt einen Überblick über die Ergebnisse der Evaluation bzgl. der Fehlerrate auf Zeichenebene unter Berücksichtigung der Vorverarbeitung des Trainings- und Testmaterials, der Modellklasse (standard vs. spezifisch) und der eingesetzten OCR-Software (OCRopus, Tesseract). Das beste (grün) und das schlechteste Ergebnis (rot) sind hervorgehoben. Da wir keinen Einfluss auf die Vorverarbeitung der Trainingsmaterialien der mitgelieferten Modelle haben, ist die Matrix in dieser Hinsicht unvollständig.

Vorverarbeitung		OCRopus		Tesseract	
Training	Test	standard	spezifisch	standard	spezifisch
nlbin	nlbin	25,41 %	6,04 %	-	53,10 %
	Scantailor	21,05 %	3,89 %	-	40,91 %
Scantailor	nlbin	-	6,95 %	37,37 %*	29,81 %
	Scantailor	-	4,21 %	27,15 %*	16,48 %

***Tab. 1**: Darstellung der Ergebnisse auf Einzel-OCR-Ebene im Bezug auf Vorverarbeitungsmethode für Trainings- und Testmaterial, Modelltyp und verwendete OCR-Software.*

Die geringste erreichte Fehlerrate (3,89 %) liegt etwa im Bereich der Textgenauigkeit der 111 Gedichte aus der Pilotstudie von Federbusch (Federbusch / Polzin 2013). Die Fehlerrate von Tesseract ist jeweils höher als die von OCRopus. Der sequenzorientierte Ansatz hat klare Vorteile bei der Erkennung von Schriftzeichen, die die typischen Charakteristika früher Drucke aufweisen.

Desweiteren zeigt sich, dass die Vorverarbeitung mit nlbin für Tesseract sowohl auf Trainings- als auch auf Testebene jeweils schlechtere Ergebnisse bringt. Für OCRopus sind die Ergebnisse bzgl. der Vorverarbeitung differenzierter: Die beste Kombination liefert eine Vorverarbeitung des Trainingsmaterials mit nlbin bei einer nachfolgenden Vorverarbeitung des Testmaterials mit Scantailor. Unterschiede im Ergebnis der Vorverarbeitung beider Programme illustriert Abbildung 3.

Abb. 3: Bild einer Textzeile nach der Vorverarbeitung mit nlbin (oben) und Scantailor (unten).

Die von Scantailor durchgeführte Bildvorverarbeitung ist deutlich normativer und für einen zeichenorientierten Ansatz wie Tesseract besser geeignet. Das Training sequenzorientierter Ansätze leidet unter dieser Vergröberung.

Es zeigt sich erneut, dass spezifisch trainierte Modelle eine massive Textgenauigkeitsverbesserung mit sich bringen können (vgl. auch Springmann et al. 2015).

Textvereinigung

Betrachtet man die Beispielausgaben in Abbildung 2, so wird der Qualitätsunterschied zwischen beiden OCR-Programmen ersichtlich. An einzelnen Stellen jedoch (z. B. Großbuchstaben am Anfang der Zeile im letzten Abschnitt) hat Tesseract Erkennungsvorteile.

Ausgehend von diesem Befund wurde der jeweils genaueste Text von OCRopus und Tesseract miteinander vereinigt. Es hat sich gezeigt, dass die Konfidenzen, die die Programme für jedes Zeichen zurückliefern, kein verlässliches Kriterium sind, um Konflikte aufzulösen. Die Fehlerrate nimmt zu. Die Strategie, Wörter bzw. Sequenzen zu bevorzugen, die sich im Referenzlexikon befinden, hat dagegen eine messbare Verbesserung mit sich gebracht. Die Anzahl der falsch erkannten Zeichen konnte um 14 % reduziert werden (Fehlerrate 3,34 %). Es ist zu vermuten, dass der Effekt größer wäre, wenn zwei OCR-Ergebnisse mit vergleichbarer Qualität vorlägen. Dies bleibt jedoch zum jetzigen Zeitpunkt für Drucke des 17. Jahrhunderts ein Desiderat.

Fußnoten

1. Verzeichnis der im deutschen Sprachraum erschienenen Drucke des 17. Jahrhunderts.
2. Vgl auch Dach (o. J.) in http://www.zeno.org/Literatur/M/Dach,+Simon/Gedichte sowie TextGrid (2015).
3. „Ziesemers Dach-Ausgabe ist textlich zu wenig genau, um auch für die dort abgedruckten, fast ausnahmslos deutschsprachigen, Gedichte den Rückgriff auf die kasualen Einzeldrucke und andere zeitgenössische Ausgaben entbehren zu können. Jede Stichprobe erweist für jedes einzelne Gedicht Transkriptionsfehler und unerklärte Texteingriffe.“ (Walter 2008: 466)
5. Für Frakturdrucke des 19. Jahrhunderts ist ein solch starker Unterschied zwischen den Tesseract und OCRopus nicht nachgewiesen.

Bibliographie

Bloomberg, Dan (2001-2015): Leptonica http://www.leptonica.com/ [letzter Zugriff: 15. Oktober 2015].

Dach, Simon (o. J.): *Gedichte* http://www.zeno.org/Literatur/M/Dach,+Simon/Gedichte [letzter Zugriff 15. Oktober 2015].

Federbusch, Maria / Polzin, Christian (2013): *Volltext via OCR - Möglichkeiten und Grenzen*. Testszenarien zu den Funeralschriften der Staatsbibliothek zu Berlin - Preußischer Kulturbesitz. Berlin Staatsbibliothek zu Berlin http://staatsbibliothek-berlin.de/fileadmin/user_upload/zentrale_Seiten/historische_drucke/pdf/SBB_OCR_STUDIE_WEBVERSION_Final.pdf [letzter Zugriff 15. Oktober 2015].

Garber, Klaus (2001-2013): *Handbuch des personalen Gelegenheitsschrifttums in europäischen Bibliotheken und Archiven*. 13 Bände. Hildesheim / Zürich / New York: Olms / Weidmann.

GitHub Inc. (2016a): *ScanTailor* http://scantailor.org/ [letzter Zugriff 15. Oktober 2015].

GitHub Inc. (2016b): *OCRopus* https://github.com/tmbdev/ocropy [letzter Zugriff 15. Oktober 2015].

GitHub Inc. (2016c): *Tesseract* https://github.com/tesseract-ocr [letzter Zugriff 15. Oktober 2015].

HAB = Herzog August Bibliothek Wolfenbüttel (2007-2016): *VD17*. Das Verzeichnis der im deutschen Sprachraum erschienenen Druck des 17. Jahrhunderts http://www.vd17.de/index.php?category_id=1&article_id=1&clang=0 .

Hunt, James W. / McIlroy, M. Douglas (1976): "An Algorithm for Differential File Comparison" in: *Computing Science Technical Report* (Bell Laboratories) 41 http://www.cs.dartmouth.edu/~doug/diff.pdf

Klöker, Martin (2010): "Das Testfeld der Poesie. Empirische Betrachtungen aus dem Osnabrücker Projekt zur 'Erfassung und Erschließung von personalen Gelegenheitsgedichten'", in: Keller, Andreas / Lösel, Elke / Wels, Ulrike / Wels, Volkhard (eds.): *Theorie und Praxis der Kasualdichtung in der Frühen Neuzeit* (= Chloe. Beihefte zu Daphne 43). Amsterdam / New York: Rodopi 39-84.

Python Software Fundation (1990-2016): *difflib - Helpers for Computing Deltas* https://docs.python.org/2/library/difflib.html [letzter Zugriff 15. Oktober 2015].

Segebrecht, Wulf (1977): *Das Gelegenheitsgedicht*. Ein Beitrag zur Geschichte und Poetik der deutschen Lyrik. Suttgart: Metzler.

Springmann, Uwe / Lüdeling, Anke / Schremmer, Felix (2015): "Zur OCR frühneuzeitlicher Drucke am Beispiel des RIDGES-Korpus von Kräutertexten (Poster)", in: *Tagung der DHd (Digitale Geisteswissenschaften im deutschsprachigen Raum)*, Graz https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/mitarbeiter-innen/anke/pdf/SpringmannLuedelingSchremmer2015.pdf [letzter Zugriff 15. Oktober 2015].

TextGrid (2015): *Die digitale Bibliothek bei TextGrid* https://textgrid.de/digitale-bibliothek [letzter Zugriff 15. Oktober 2015]

VietOCR http://vietocr.sourceforge.net/ [letzter Zugriff: 15. Oktober 2015].

Vobl, Thorsten / Gotscharek, Annette / Reffle, Uli / Ringlstetter, Christoph / Schulz, Klaus U. (2014): "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 (DATeCH '14)*: 57-61 http://dl.acm.org/citation.cfm?id=2595197 [letzter Zugriff 15. Oktober 2015].

Walter, Axel E.(2008): "Dach digital? Vorschläge zu einer Bibliographie und Edition des Gesamtwerks von Simon Dach nebst einigen erläuterten Beispielen vernachlässigter bzw. unbekannter Gedichte", in: Walter, Axel E. (ed.) in: *Simon Dach (1605–1659)*. Werk und Nachwirken. Tübingen: Niemeyer: 465-522.

Ziesemer, Walter (ed.) (1936-1938): *Simon Dach: Gedichte*. Vier Bände. Halle an der Saale: Niemeyer.