Verbesserung der OCR in digitalen Sammlungen von Bibliotheken

Konstantin Baierer[1] und Philipp Zumstein[2]

[1] *Universitätsbibliothek Mannheim,*
konstantin.baierer@bib.uni-mannheim.de
[2] *Universitätsbibliothek Mannheim,*
philipp.zumstein@bib.uni-mannheim.de

Abstract

Möglichkeiten zur Verbesserung der automatischen Texterkennung (OCR) in digitalen Sammlungen insbesondere durch computerlinguistische Methoden werden beschrieben und bisherige PostOCR-Verfahren analysiert. Im Gegensatz zu diesen Möglichkeiten aus der Forschung oder aus einzelnen Projekten unterscheidet sich die momentane Anwendung von OCR in der Bibliothekspraxis wesentlich und nutzt das Potential nur teilweise aus.

Possibilities for improving the optical character recognition (OCR) in digital collections in particular by using computationally linguistical methods are described and previous PostOCR-techniques are analyzed. On contrast to these methods from the research and single projects, the current use of OCR in library practice differs essentially and does use its potential only partially.

1 Einleitung

Bibliotheken bieten seit jeher Zugang zu ihrem Bestand und dies hat sich auch im Informationszeitalter im Prinzip nicht geändert. Der Zugang zum gedruckten Bestand ist dabei an den Ort gebunden im Gegensatz zum elektronischen Bestand. Durch die Digitalisierung können historisch wertvolle und für die Forschung besonders wichtige Bibliotheksbestände auch online und damit ortsungebunden zur Verfügung gestellt werden. Die digitalisierten Bestände werden in sogenannten digitalen Sammlungen von der Einrichtung selbst und häufig zusätzlich über Aggregatoren (z.B. Europeana, DDB, e-rara) zugänglich gemacht. Neben den reinen Bild-Digitalisaten und bibliografischen Metadaten werden auch Auszeichnungen wie etwa die Grobstruktur auf Kapitelebene oder Paginierung vorgenommen.

Die Strukturdatenerfassung und Online-Zugänglichkeit vereinfacht die Recherche für Forschende bereits erheblich, aber erst die Durchsuch- und Analysierbarkeit der vollständigen Werke auf Basis des Volltextes macht aus „digitalem Mikrofilm“ wertvolle Forschungsdaten. Statt den teuren und aufwändigen Verfahren der manuellen Transkription, insbesondere im Double-Key-Verfahren, wird vermehrt Software für die Texterkennung (OCR) verwendet. Wie bei vielen Automatismen ist das Ergebnis fehlerbehaftet, wobei zu beachten ist, dass eine 100% Erkennungsgenauigkeit auch mit manuellen Transkriptionen normalerweise nicht

erreicht wird, z.B. macht eine Schreibkraft in 100 Wörtern durchschnittlich 8 Tippfehler (Ratatype 2013). Verschiedene Forschungsmethoden (insbesondere mit Hilfe von Information Retrieval) können zudem auch mit fehlerbehafteten Volltexten umgehen. Sehr fehlertolerant ist auch der Mensch, welcher häufig einen Text mit ein paar Fehlern ohne Probleme lesen und verstehen kann, und wahrscheinlich haben sich sogar ein paar 3ehler in in diesen Text eingeschljchcn:-)

Wir werden zunächst die theoretischen Grundlagen zur Verbesserung der OCR erläutern und im zweiten Teil auf die Praxis in Bibliotheken zu sprechen kommen.

2 OCR und Erkennungsgenauigkeit

Bei der OCR werden Methoden aus der Bildverarbeitung, künstlichen Intelligenz (etwa neuronale Netze) und auch Computerlinguistik verwendet. Der OCR-Prozess ist schematisch in Abb. 1 dargestellt, wobei dabei bewusst auch etwas verallgemeinert wurde. Die kommerziellen OCR-Lösungen von ABBYY beinhalten bereits einen Wörterbuchabgleich, welcher durch die Angabe einer Sprache gesteuert wird. Die Open-Source-Lösung Tesseract verfügt auch über Sprachmodelle, gewichtet sie aber geringer. Dahingegen fokussiert sich die ebenfalls freie Software Ocropus auf den Bildverarbeitungsprozess mit neuronalen Netzen und verzichtet auf spezifische Sprachmodelle. Daher unterscheiden sich die Ergebnisse dieser OCR-Programme meist substantiell, insbesondere in der Art der potentiellen Erkennungsfehler: Falls der Wörterbuchabgleich während der Texterkennung bereits relativ stark gewichtet wird, muss vermehrt mit „Verschlimmbesserungen“ gerechnet werden, d.h. der Ersetzung von korrekt erkannten aber nicht im Wörterbuch verzeichneten durch sprachlich richtige aber nicht der Vorlage entsprechenden Wörtern, vor allem bei unerwarteten Eigennamen, Sprachen, Schriften, Schreibweisen.

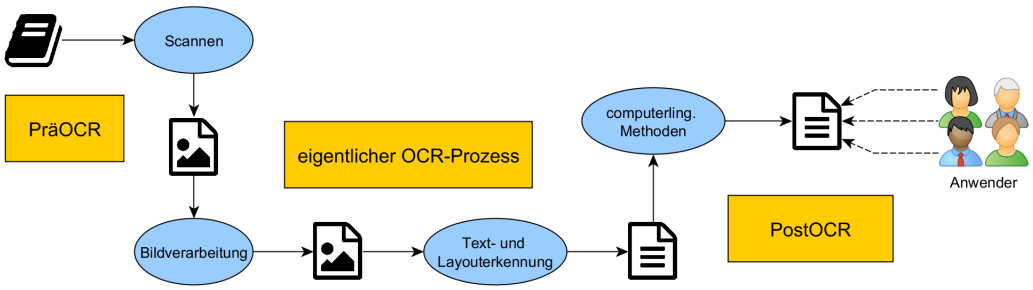

Abbildung 1: Die verschiedenen Module des OCR-Prozesses inkl. der Phasen davor (PräOCR) und danach (PostOCR) sind hier schematisch benannt: Scannen, Bildverarbeitung, Text- und Layouterkennung, computerlinguistische Methoden. Dabei findet jeweils wie angedeutet eine Medientransformation vom gedruckten Buch über das Bild-Digitalisat zum erkannten Text statt. Anwender können dann das Ergebnis für unterschiedliche Zwecke nutzen

Brauchbare OCR für Positiv-Suchen sollte nach den DFG-Praxisregeln „Digitalisierung“ mindestens 80% Buchstabengenauigkeit erreichen, ab 99,95% gilt der Text auch für ausschliessende Suchen als zuverlässig. Alles dazwischen wird teilweise auch als „schmutzige OCR“ bezeichnet, insbesondere falls die automatische Texterkennung nicht weiter nachkorrigiert wird. Auch wenn dieser Fokus auf Zeichengenauigkeit und Information Retrieval die

Probleme in der Praxis nicht immer widerspiegelt, so ist doch die Verbesserung der Erkennungsgenauigkeit ein hohes Ziel für digitale Sammlungen, um die optimale Nachnutzbarkeit von Volltexten zu ermöglichen.

Mit den gängigen Off-the-shelf OCR-Lösungen können hohe Zeichengenauigkeiten, die den Anforderungen der DFG gerecht werden, lediglich bei gedruckten, einfach layouteten Werken mit modernem Antiqua-Schriftsatz in natürlicher Sprache mit heute üblicher Grammatik und Typografie und bei Vorhandensein eines trainierten Modells für diese Sprache erreicht werden, nicht aber für

- Gedruckte Werke vor 1945, insbesondere Fraktur
- Zeitungen und andere Werke mit nicht-trivialem Layout und Lesefluss
- Überwiegend nicht-natürlichsprachliche Texte (bspw. Statistiken)
- Texte mit vielen Eigennamen (bspw. „Who is Who“, Verwaltung, Landkarten, Kochbücher...)
- Texte mit typografischen Idiosynkrasien (bspw. „langes s“ ſ in Antiqua-Schrift)
- Mehrere Sprachen/Schrifttypen pro Werk

Wenn man die eigentliche OCR-Erkennung nicht weiter ändern möchte oder kann und die PräOCR-Schritte bereits optimiert sind, bietet sich zur Erhöhung der Zeichengenauigkeit nur an, dies im Nachgang, d.h. in der PostOCR-Phase, anzugehen. Häufig wird dabei der Fokus auf den erkannten Text gelegt und mit Wörterbüchern gearbeitet um möglichst viele Fehler zu verbessern, wie wir weiter unten noch ausführen werden. Solche Ansätze behandeln nur einige der Probleme in den oben aufgeführten Punkten und können auch nicht alle Arten von Fehlern korrigieren.

Die unterschiedlichen Fehler, welche bei der OCR auftreten können, unterscheiden sich von Tippfehlern erheblich. Nicht alle Fehler sind für einen bestimmten Anwendungsfall gleich relevant. Daher ist es auch immer nützlich die Fehler genauer zu analysieren, um geeignete Methoden zu wählen.

Eine Kategorisierung dieser OCR-Fehler ist eine natürliche Herangehensweise zur weiteren Analyse und kann als Grundlage für mögliche Korrektur-Schritte in einem PostOCR-Verfahren dienen. Wir möchten hierbei vier Hauptkategorien unterscheiden: Zeichenfehler, Abstandsfehler, Wortfehler, Leseflussfehler. Jede Hauptkategorie kann dann wieder weitere spezifischere Fehlerkategorien enthalten:

1. Zeichenfehler
    (a) Falsch erkannte Zeichen (z.B. anstatt e wird c erkannt oder rn wird als m erkannt)
    (b) Falsch erkannte Akzente, Umlaut-Auszeichnungen (z.B. e wird anstatt é oder ê erkannt)
    (c) Zeichen-Zahlen-Vertauschungen (z.B. I und 1 werden vertauscht)
    (d) Zeichenverdopplungen
    (e) Falsch erkannte Interpunktionszeichen, Währungszeichen o. ä.

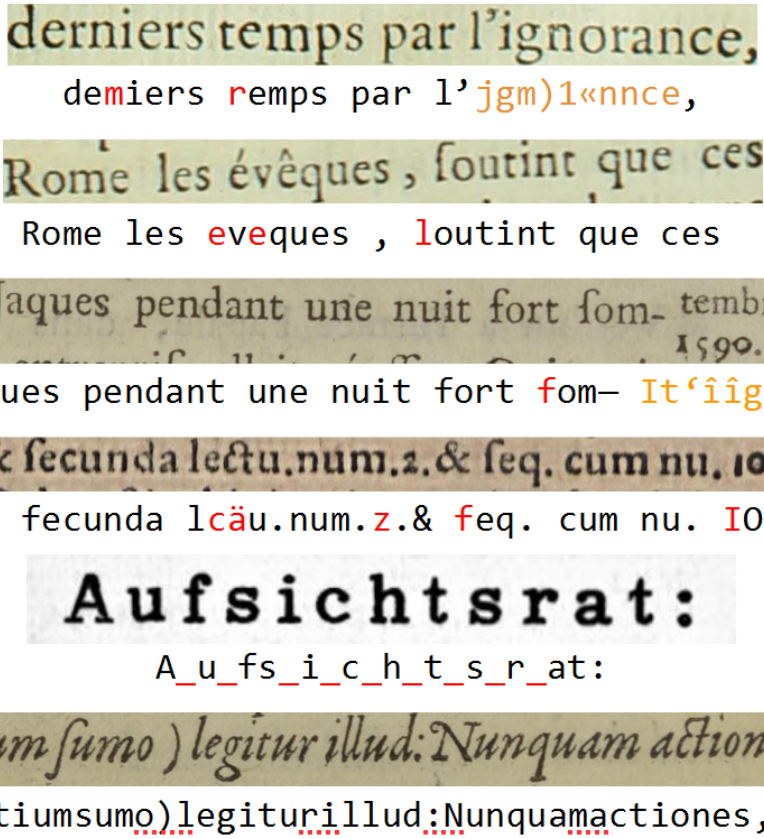

Abbildung 2: Beispielbilder von Zeilen aus Digitalisaten und dem erkannten Text einer OCR. Die Fehler sind jeweils farblich hervorgehoben (entweder zeichenweise, wortweise oder bei den Abständen)

2. Abstandsfehler
   (a) Segmentierungsfehler (Wortsplits etwa durch gesperrte Schrift oder Verschmelzungen von verschiedenen Wörtern durch einen nicht erkannten Abstand)
   (b) Abstände unterschiedlich erkannt (Abstände vor oder nach Interpunktationszeichen, mehrere Abstände)
3. Wortfehler
   (a) Wörter werden komplett falsch erkannt (z.B. beim Wechsel der Schriftart)
   (b) Text wird in Schmutz- oder Bilderbereichen erkannt
   (c) Falsch erkannte Zeilen durch spezielles Layout (z.B. zwei Zeilen werden als eine erkannt und dann versucht darin Text zu erkennen)
   (d) „Verschlimmbesserungen“
4. Leseflussfehler
   (a) verschiedene Spalten werden in einem Block wiedergegeben oder Marginalien dem Fliesstext zugeschlagen
   (b) komplexe Layoutfehler etwa beim Lesefluss in Zeitungen

Andere Klassifikationen von OCR-Fehlern findet man bei Niklas (2010:4–5) sowie Wernersson (2015), aus der zweiten Quelle haben wir die Kategorie „Leseflussfehler“ übernommen. Die verschiedenen Fehler sind dann jeweils immer in Hinblick auf die Anwendung zu beurteilen: „Um (sinnvolle) Aussagen über die Gesamtqualität der OCR-Daten treffen zu können, muss man berücksichtigen, wie die OCR-Daten später genutzt werden sollen. Die Daten sind ohne die Angabe, was mit ihnen erreicht werden soll, nicht zu beurteilen.“ (Wernersson 2015:34) Beispielsweise werden bei einer Stichwort-Suche heutzutage häufig auch verschiedenen Normalisierungen mitgesucht, so dass die Akzente meist nebensächlich sind, ebenfalls dürften etwa Interpunktionszeichen und Abstände, welche nicht zu Segmentierungsfehlern führen, keinen Einfluss auf eine Recherche nach Stichwörtern haben. Auf der anderen Seite beeinflussen Segmentierungsfehler die Möglichkeiten zur Recherche nicht unwesentlich und „[d]ie Wortsegmentierung ist eines dieser Probleme, das bisher noch im Schatten anderer OCR-Textprobleme verborgen lag“ (Shashkina 2010:5). Auch werden etwa Zeichen-Zahlen-Vertauschungen die Suche nach Stichwörtern oder Extraktion von bestimmten Zahlen erschweren bzw. gar verunmöglichen.

3 Computerlinguistische Methoden auf OCR-Texten

Eine bekannte Anwendung der Computerlinguistik sind Korrektursysteme (im Englischen meist einfach als „spell checking“ bezeichnet), welche in gängigen Schreibprogrammen als Rechtschreibprüfung und -korrektur implementiert sind. Dabei werden sprachabhängig Wörterbücher, Grammatikregeln und weitere Modelle herangezogen, um falsch geschriebene Wörter und Sätze zu markieren bzw. Korrekturvorschläge zu geben.

Prinzipiell kann man fehlerhaft erkannten Text einer OCR als falsch geschriebenen Text interpretieren und in der PostOCR-Phase die gleichen oder ähnliche computerlinguistische Methoden anwenden, um einen Text mit weniger Fehlern herauszubekommen.

Für einen Überblick über Korrektursysteme in der Computerlinguistik empfehlen wir Fliedner (2010) sowie Mitton (2010). Darin wird die Korrektur von Nichtwörtern klar von der kontextabhängigen Korrektur unterschieden. Die Korrektur von Nichtwörtern erfolgt dabei meist als eine Art Wörterbuchabgleich, was wir im Folgenden noch weiter beleuchten. Bei der kontextabhängigen Korrektur wird angenommen, dass zwar jedes Wort für sich alleine ein gültiges Wort ist, sich aber im Kontext Fehler ergeben können. Es wird berichtet, dass ca. 40% der Tippfehler kontextuelle Fehler sind (Kukich 1992), aber dies dürfte bei den OCR-Fehlern sehr viel niedriger sein. Die Erkennung solcher Fehler muss daher aus dem Kontext geschehen und kann sich dabei beispielsweise auf die häufigen Wortverwechslungen (Confusion sets) und Wort-n-Gramme (z.B. n=3) stützen. Auch weitere Prüfungen wie etwa eine Grammatikprüfung sind in solchen Korrektursystemen möglich.

Beim Vergleich von Tippfehlern und OCR-Fehlern sowie Methoden zu deren Erkennung bzw. Korrektur gibt es neben allen Gemeinsamkeiten auch grössere Unterschiede wie etwa:

(a) Tippfehler sehen wesentlich anders aus als die üblichen Fehler bei einer OCR.

(b) Alte Schreibweisen sollen bei einer genauen Transkription erhalten bleiben und nicht an die neue Schreibweisen angepasst werden.

4 Analyse bisheriger PostOCR-Verfahren

Bisherige PostOCR-Verfahren setzen als zentrale Komponenten meist auf einen Wörterbuchabgleich. Als ersten Schritt wird dabei das Dokument der OCR-Ausgabe in die einzelnen Wörter zerlegt, häufig wird auch einfach die von der OCR-Software gefundene Aufteilung übernommen. Danach wird jedes Wort s in einem Wörterbuch nachgeschlagen. Falls s im Wörterbuch enthalten ist, dann ist keine Korrektur notwendig. Andernfalls kann das Nichtvorhandensein von s im Wörterbuch auf einen möglichen Fehler hindeuten, da im Wörterbuch keine falsch geschriebenen Wörter vorkommen. Dann kann man versuchen zu s mögliche Ersetzungskandidaten aus dem Wörterbuch zu berechnen. Um einen Kandidaten auszuwählen, welcher am nächsten zu s liegt, braucht man ein Ähnlichkeitsmass oder vergleichbare Methoden.

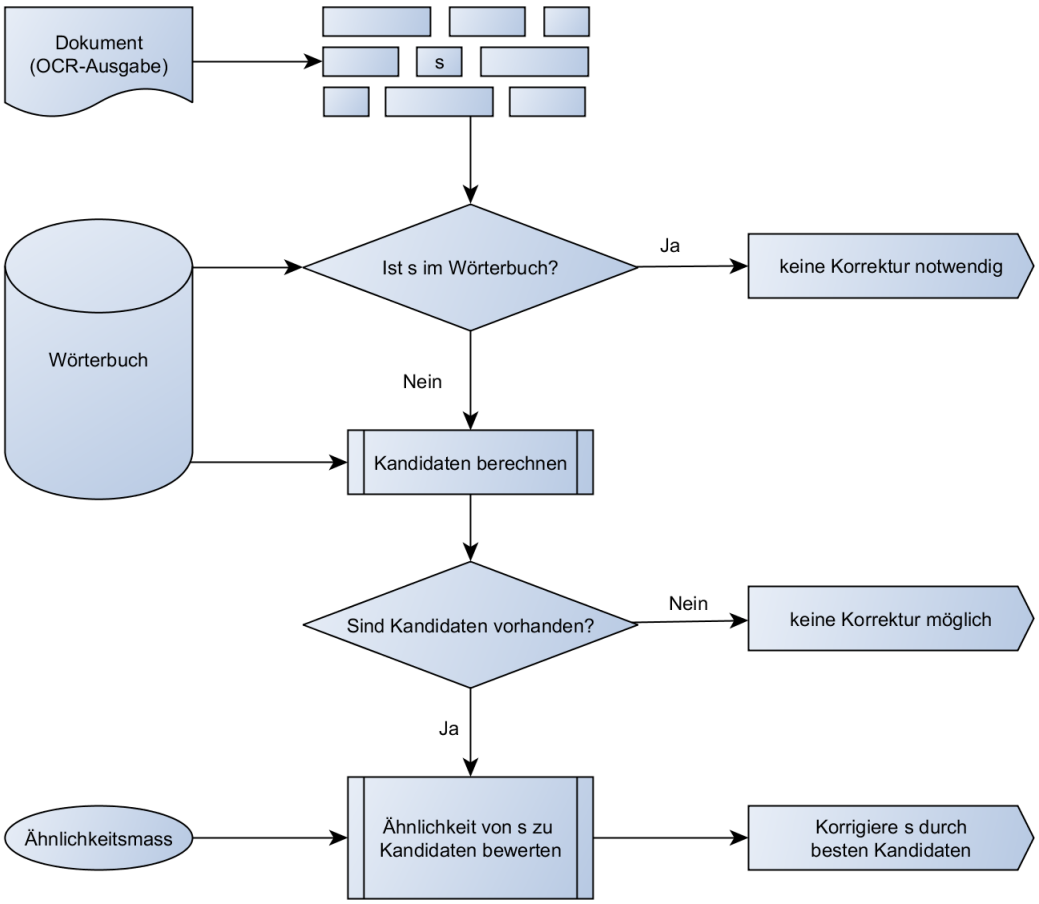

Abbildung 3: Wörterbuchabgleich schematisch: Einzelne Implementationen unterscheiden sich dann im verwendeten Wörterbuch bzw. dessen Erstellung, der Berechnung von Kandidaten und dem verwendeten Ähnlichkeitsmass sowie weiteren möglichen Regeln

4.1 Wörterbuch und Sprachmodelle

Das Ziel eines Wörterbuches ist es die Wörter einer Sprache zu verzeichnen und die richtige Schreibweise anzugeben. Ein Autor kann daher beim Schreiben eines Textes die korrekte

Schreibweise eines einzelnen Wortes nachschlagen. Ähnlich kann man ein erkanntes Wort in einem OCR-Dokument in einem PostOCR-Schritt in einem Wörterbuch nachschlagen. Neben dem eigentlichen Wort sind dabei häufig auch Deklinations- oder Konjugationsformen beschrieben. In einem elektronischen Wörterbuch entspricht dies dann meistens Morphologie-Regeln zur Agglutination/Flexion, Derivation bzw. Komposition (Fliedner 2010:557).

Der einfachste Ansatz ist die Verwendung eines fixen vorgegebenen Wörterbuchs. Häufig genutzte und freie Wörterbücher von Korrektursystemen findet man dabei etwa bei ispell[1], OpenThesaurus[2], Wiktionary[3]. Die meisten Sprachen sind damit in der heutigen geltenden Schreibweise abgedeckt, wobei die Qualität und Abdeckung zwischen verschiedenen Sprachen unterschiedlich sein kann. Bei historischen Texten helfen diese Wörterbücher aber nur beschränkt weiter, da die alten Schreibweisen ja keine OCR-Fehler sind, sondern buchstabengetreu wiedergegeben werden.

In Nölte u. a. (2016) wurde ein angepasstes Wörterbuch durch eine Liste historischer bzw. dialekt- oder fachspezifischer Wortformen aus einem entsprechenden Korpus vom Deutschen Textarchiv (DTA) erstellt. Die Werke im DTA stehen im Volltext zur Verfügung, welcher grösstenteils mittels Double Key, experimentell durch OCR und mit nachträglicher manueller Nachkorrektur erstellt wurde und somit hohen Qualitätsansprüchen genügt. Die Wörter in diesem Korpus können als neues Wörterbuch für historische Schreibweisen zusammengefasst werden. Auf die gleiche Weise kann man ausgehend von anderen Korpora mit relevanten, qualitativ hochwertigen Volltexten ein angepasstes Wörterbuch erzeugen.

Bestehende Wörterbücher kann man auch manuell anreichern, bspw. aus einzelnen manuell korrigierten Zeilen („Ground Truth“, die auch für die Qualitätsabschätzung benutzt werden kann) oder aus manuell eingegebenen Kapitelüberschriften (sowie diese bei der Strukturdatenerfassung eingegeben werden). Im Projekt zur OCR von Funeralschriften (Federbusch 2015) wurde dies bereits erprobt und vorgeschlagen, kooperativ die bereits vielerorts erfassten Strukturdaten auch für Wortlisten nutzbar zu machen.

Man kann auch aus dem fehlerbehafteten OCR-Text die häufigsten Wörter heraussuchen und als Wörterbuch für den Rest betrachten. Falls ein Wort häufig auftritt, ist es relativ unwahrscheinlich, dass dieses immer (gleich) falsch erkannt worden ist. Ebenfalls kann man sich hier natürlich auch ein semi-automatisches Verfahren vorstellen, bei dem nur Vorschläge für Wörterbucheinträge berechnet werden und diese noch von einem Bearbeiter bestätigt werden müssen.

4.2 Kandidatensuche und Ähnlichkeitsmasse

Als Standardmass zwischen zwei Zeichenketten hat sich die Levenshtein-Distanz (auch „edit distance“) eingebürgert. Dabei wird die minimale Anzahl atomarer Operationen gezählt, um eine Zeichenkette in eine andere umzuwandeln. Eine atomare Operation ist dabei das Einfügen, Löschen oder Ändern eines einzelnen Zeichens. Teilweise wird auch das Vertauschen von zwei Zeichen als eine Grundoperation gezählt (Damerau-Levenshtein-Distanz[4]), was im

[1] http://fmg-www.cs.ucla.edu/geoff/ispell-dictionaries.html
[2] https://www.openthesaurus.de/about/download
[3] https://dumps.wikimedia.org/dewiktionary/latest/
[4] https://de.wikipedia.org/w/index.php?title=Levenshtein-Distanz&oldid=155493707

Hinblick auf Tippfehler auch gut verständlich ist. Suche nach Worten mit konfigurierbarer Levenshtein-Distanz wird von den grossen Suchmaschinen-Softwarebibliotheken (Lucene, Solr, ElasticSearch) ebenso unterstützt wie vom Standard-Linux-Tool agrep.

Nun kann man diese Grundoperationen aber auch unterschiedlich gewichten. Für PostOCR-Verfahren liegt es nahe die typischen OCR-Fehler zu berücksichtigen (Nölte u. a. 2016; Evershed und Fitch 2014). Beispielsweise sollten die beiden Wörter „echt“ und „ccht“ näher beieinander liegen als „echt“ und „mcht“, da die Verwechslung von e, c für eine OCR wesentlich plausibler ist, als eine Verwechslung von e, m. Neben einzelnen Zeichen können in solchen Verwechslungen auch mehrere Zeichen durch ein oder mehrere Zeichen ersetzt werden. Diese üblichen Verwechslungen/Konfusionen („confusions“) kann man durch die Analyse der OCR-Fehler in einem manuell korrigierten Referenztext berechnen.

Diese Ähnlichkeitsmasse kann man für die Berechnung von Kandidaten verwenden oder auch um unterschiedliche Kandidaten zu vergleichen und den bestmöglichen auszuwählen. Im Gegensatz dazu kann man die Modellierung auch mit einem Wahrscheinlichkeitsmodell „Noisy-Channel-Modell“ machen und die OCR-Fehler etwa als Folge von zufälligen Verrauschungen betrachten. Dabei können den oben genannten Konfusionen entsprechend Wahrscheinlichkeiten zugeordnet werden. Die PostOCR kann dabei als Prozess verstanden werden, mit dem Ziel den wahrscheinlichsten und möglichst fehlerfreien Text zu berechnen (Tong und Evans 1996).

Eine andere Methode, um die Kandidaten zu einem Wort zu berechnen, benutzt das Anagramm-Hashing (Reynaert 2010). Dabei werden Wörter, welche aus denselben Buchstaben bestehen, aber nicht zwingend in derselben Reihenfolge, auf die gleichen Hashwerte abgebildet. Der Hashwert von falsch erkannten Wörtern kann dann mit dem Hashwert der Wörter im Wörterbuch verglichen werden, um mögliche Kandidaten für die Korrektur zu bekommen.

Ein anderer Hashing-Ansatz mit dem Namen OCR-Key wird bei Niklas (2010) beschrieben. Jedem Wort wird ein Ähnlichkeitsschlüssel zugeordnet und Wörter mit dem gleichen Schlüsselwert gelten dann als mögliche Kandidaten. Die Besonderheit hierbei ist die Analyse der Buchstaben und Buchstabenkombinationen bzgl. Merkmale, welche auch bei Verunreinigungen in Digitalisaten oder schlechtem Druck möglichst erhalten bleiben. Beispielsweise gehören iii und m zur gleichen Klasse, charakterisiert durch drei vertikale Striche.

Bei Evershed und Fitch (2014) wird zusätzlich eine Reversed OCR Prozedur beschrieben, bei welchem die Ähnlichkeit zwischen zwei Textstrings aus der Vergleichsanalyse der generierten Bilder dieser Wörter in einer geeigneten Schriftart berechnet wird.

Prinzipiell kann man als Kandidatenmenge auch alle Wörter im Wörterbuch betrachten, aber dies würde häufig zu rechenintensiv werden. Eine eingeschränkte Kandidatenmenge wird bei Tong und Evans (1996) durch Betrachtung der n-Grams der Buchstaben innerhalb eines Wortes (inklusive Worttrennungszeichen) beschrieben.

5 OCR-Verbesserung in der Praxis

Es mangelt also nicht an guten und kontinuierlich weiterentwickelten Ideen, wie mit dem Wissen um die Musterhaftigkeit von Sprache und Typografie in Kombination mit neuen Methoden und Datensets den OCR-Fehlern der Garaus gemacht werden kann. Betrachtet man im weiteren Kontext die Forschungsergebnisse und technischen Neuerungen der letzten zwei Jahrzehnte, zeigt sich, dass OCR und verwandte Techniken enorme Fortschritte gemacht haben und ein hohes Mass an Professionalisierung erreicht haben, durch effektivere Algorithmen (bspw. neuronale Netze statt Markov-Modelle), cloudbasierte Infrastruktur (bspw. reCaptcha, OCR SDK), riesige für Training verfügbare Datenmengen und Evaluation (bspw. Wikisource, Google Books).

Tatsächlich erreichen viele der veröffentlichten Ansätze Verbesserungen auf besonders schwierigen Daten und mit besonders effektivem Ansatz auch 2016 noch beeindruckende Verbesserungen der Erkennungsgenauigkeit. Gleichwohl ist die Forschung in diesem Bereich inzwischen so ausdifferenziert, dass in den diversen Competitions um einzelne Prozentpunkte in der Erkennungsgenauigkeit gerungen wird und das auch nur für Werke mit komplexen Layouts oder handschriftlichem Text. Für die Forschungsabteilung von Google etwa ist OCR ein im Grunde gelöstes Problem (Popat 2015). Gleichzeitig wächst die Kluft zwischen Forschungsstand und Implementierung seit Jahren kontinuierlich (Nagy 2015).

In der Bibliothekswelt sind die Digitalen Sammlungen, die zusätzlich zu den Bild-Digitalisaten auch hochwertige Volltexte bereitstellen, in der absoluten Minderheit. Die Historiker aller Couleur, die Digital Scholars, Data Mining Experten und Korpuslinguisten wissen oft gar nicht, dass es Schätze in VD16/17/18, der Deutschen Digitalen Bibliothek und anderen digitalen Sammlungen gibt, und wenn sie es wissen, können Sie ihre textbasierten Methoden nicht auf die Bild-Digitalisate anwenden.

Der digitale Wandel hat sein disruptives Potential voll entfaltet und viele der Anwendungsfälle obsolet gemacht, die OCR für kommerzielle Hersteller interessant machten: Die Erkennung von Formularen, Patenten, Rechnungen und ähnlich rigide strukturierten Dokumenten ist eine endliche Aufgabe, da papierbasierte Kommunikation zurückgeht (Nagy 2015). Die starke Konsolidierung des Markts für OCR-bezogene Software, sowie die geringe kommerzielle Viabilität von Digitalisierung von Kulturgut haben dazu geführt, dass für viele Bibliotheken OCR und der Name der marktbeherrschenden Firma gleichbedeutend sind und automatisierte Volltexterfassung eine Black Box ist, in die Geld und Bild-Digitalisate gegeben werden und aus der Volltextergebnisse herauskommen.

Die Ergebnisse werden meist mit der DFG-Methode für Zeichengenauigkeit stichprobenartig evaluiert. Es stellt sich dann oft als schwerer heraus als gedacht oder wird gar nicht versucht, die verbleibenden Fehler noch zu verbessern, insbesondere bei scheinbar bereits sehr guten Ergebnissen (bspw. > 97%). Im Haus entwickelte ad hoc Software-Lösungen werden häufig nicht weiterentwickelt, dokumentiert oder veröffentlicht, weil sie zu spezifisch scheinen und OCR als aufwändiger aber einmaliger Vorgang wahrgenommen wird. Wenn im Rahmen eines Projekts eine nicht erreichte Genauigkeit zugesagt wurde, könnte man nun die PostOCR-Schritte an einen anderen Dienstleister delegieren, oder, wenn nichts anderes hilft, manuell transkribieren.

Diese für alle ausser den kommerziellen Dienstleistern unbefriedigende Situation hat aus unserer Sicht drei wesentliche Gründe:

- *Vendor Lock-In*: Wenn die zentrale Komponente im OCR-Workflow alternativlos eine proprietäre Lösung ist und Bibliotheken nicht proaktiv Open-Source-Alternativen wie Tesseract oder Ocropus zumindest evaluieren, können diese auch nicht verbessert werden. Insbesondere können diese freien OCR-Engines mit der richtigen Konfiguration und Trainingsdaten, Sprachmodellen usw. die besseren Ergebnisse liefern, bleiben aber in der Handhabbarkeit hinter kommerziellen Lösungen zurück.
- *OCR als Wasserfallmodell*: Der Prozess vom gedruckten Buch zum digitalen Textdokument wird bisher als ein grosser, einmaliger Vorgang angesehen, muss aber zwangsläufig ein iterativer Prozess sein, weil hundertprozentige Genauigkeit für alle denkbaren Anwendungsfälle nie erreicht werden kann. Wie können einzelne Seiten oder auch Seitenabschnitte neu gescannt, die OCR mit einem anderen Sprachmodell darauf angewandt, mit einem speziellen Wörterbuch, Ähnlichkeitsmass oder neuen Verfahren aus der Forschung nachkorrigiert werden? Wie können Korrekturen von Anwendern eingepflegt und spezielle Forschungswünsche adressiert werden? Dafür muss es langfristige Verantwortlichkeiten, Prozesse und ggf. auch Infrastruktur geben.
- *Zu wenig Forschung und Entwicklung*: Um die Qualität der Texterfassung beurteilen zu können, reicht es häufig nicht aus, diese auf ein einzelnes Mass herunter zu brechen. Vielmehr müssen Anwendungsszenarien einbezogen werden und Strategien entwickelt werden, um effektiv zu verbessern. Dafür muss Expertise aufgebaut und Werkzeuge entwickelt werden, insbesondere über die eigene Einrichtung und das spezielle Projekt hinaus.

In jüngster Zeit ist allerdings viel Bewegung in das Thema gekommen und es gibt viele Projekte, die Hoffnung auf mehr Nachhaltigkeit im Bereich OCR machen.

6 Ausblick

Momentan lässt sich sagen, dass GitHub die Drehscheibe für die Kooperation in der OCR-Infrastrukturentwicklung ist[5], etwa für die OCR-Dateiformate ALTO und hOCR oder die freien OCR-Engines Tesseract, Ocropus und Kraken, dem OCR-as-a-Service-Projekt open-ocr und einer Vielzahl kleinerer und grösserer Projekte rund um das Thema, die aktiv kollaborativ entwickelt werden, auch von Bibliotheken.

Hoffnungsvoll stimmt ebenfalls, dass Forschungsförderer und Politik die Situation erkannt haben und im Gefolge der Förderung von massenweiser Buchdigitalisierung der 2000er Jahre nun auch OCR und die Digitalisierung von anspruchsvollerem, textuellen Kulturgut wie Zeitungen und Handschriften ausbauen. Das DFG-Projekt OCR-D[6] wird ab 2017 eine ganze Reihe von Projekten rund um OCR koordinieren, die mit Sicherheit die OCR in Bibliotheken verbessern. Das frei verfügbare PostOCR-Werkzeug PoCoTo (Post Correction Tool)[7], das die beschriebenen computerlinguistischen Methoden schon heute für Bibliotheken nutzbar macht, wird kontinuierlich weiterentwickelt. Das EU-Projekt READ[8] und seine Transkribus-

[5] https://github.com/kba/awesome-ocr
[6] http://ocr-d.de/
[7] https://github.com/cisocrgroup/PoCoTo
[8] Recognition and Enrichment of Archival Documents, 2016-2020, https://read.transkribus.eu/

Plattform, das Schweizer Projekt HisDoc2[9] oder das BMBF-Projekt Kallimachos[10] haben Potential, auch den OCR-Prozess zu bereichern und können dabei auch auf bereits entwickelte Verfahren und Werkzeuge zu OCR aus dem Projekt IMPACT (improving access to text, 2008-2012) zurückgreifen. Daneben gibt es noch eine Reihe von Projekten, die sich um die Zeitungsdigitalisierung bemühen, die für komplexes Layout, Lesefluss, Schrifttypen und Textstruktur von Zeitungen Lösungen entwickeln müssen, die auch der OCR in Bibliotheken zugutekommen können.

Die wissenschaftlichen Grundlagen für alle Phasen der OCR sind solide genug, dass Bibliotheken sich verstärkt selbst in die Evaluation und Weiterentwicklung von Dokumentation, Werkzeugen und Standards einbringen können.

Literatur

Deutsche Forschungsgemeinschaft (2013). *DFG-Praxisregeln „Digitalisierung“*. URL: http://www.dfg.de/formulare/12_151/12_151_de.pdf.

Evershed, J. und Fitch, K. (2014). Correcting Noisy OCR: Context Beats Confusion. In: *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14)*. New York, NY: ACM. DOI: 10.1145/2595188.2595200.

Federbusch, M. (2015). OCR für Drucke der Frühen Neuzeit? Erfahrungen und Perspektiven am Beispiel von Funeralschriften. In: *Bibliotheksdienst* 49.7. DOI: 10.1515/bd-2015-0082.

Fliedner, G. (2010). Korrektursysteme. In: *Computerlinguistik und Sprachtechnologie*. Hrsg. von K.-U. Carstensen. 3. Aufl. Heidelberg: Spektrum Akademischer Verlag. DOI: 10.1007/978-3-8274-2224-8_5.

Kukich, K. (1992). Techniques for Automatically Correcting Words in Text. In: *ACM Computing Surveys* 24.4. DOI: 10.1145/146370.146380.

Mitton, R. (2010). Fifty years of spellchecking. In: *Writing Systems Research* 2.1. DOI: 10.1093/wsr/wsq004.

Nagy, G. (2015). Disruptive developments in document recognition. In: *Pattern Recognition Letters* 79. DOI: 10.1016/j.patrec.2015.11.024.

Niklas, K. (2010). *Unsupervised Post-Correction of OCR Errors. Diplomarbeit*. Leibniz Universität Hannover. URL: http://tahmasebi.se/Diplomarbeit_Niklas.pdf.

Nölte, M., Bultmann, J. P., Schünemann, M. und Blenkle, M. (2016). Automatische Qualitätsverbesserung von Fraktur-Volltexten aus der Retrodigitalisierung am Beispiel der Zeitschrift Die Grenzboten. In: *o-bib. Das offene Bibliotheksjournal* 3.1. DOI: 10.5282/o-bib/2016H1S32-55.

Popat, A. C. (2015). *Optical Character Recognition for Most of the World's Languages*. URL: http://ewh.ieee.org/r6/scv/sps/20150903Abstract.html.

Ratatype (2013). *Average Typing Speed Infographic*. URL: http://www.ratatype.com/learn/average-typing-speed/.

Reynaert, M. W. C. (2010). Character confusion versus focus word-based correction of spelling and OCR variants in corpora. In: *International Journal on Document Analysis and Recognition (IJDAR)* 14.2. DOI: 10.1007/s10032-010-0133-5.

[9] Historical Document Analysis, Recognition and Retrieval, 2009-, https://diuf.unifr.ch/main/hisdoc/hisdoc2

[10] http://kallimachos.de/

Shashkina, A. (2010). *Wortsegmentierungsprobleme und Information Retrieval auf OCR-erfassten historischen Dokumenten. Masterarbeit*. Ludwig-Maximilians- Universität München. URL: http://www.cip.ifi.lmu.de/~shashkina/MA/Wortsegmentierung_IR_OCR_historische_Dokumente_master_thesis.

Tong, X. und Evans, D. A. (1996). A statistical approach to automatic OCR error correction in context. In: *Proceedings of the fourth workshop on very large corpora*. URL: http://www.aclweb.org/anthology/W/W96/W96-0108.pdf.

Wernersson, M. (2015). Evaluation von automatisch erzeugten OCR-Daten am Beispiel der Allgemeinen Zeitung. In: *ABI Technik* 35.1. DOI: 10.1515/abitech-2015-0014.