Erfahrungsbericht zum Projekt
„Nachbearbeitung des OCR-Volltextes der Zeitschrift Die Grenzboten“
DFG-Geschäftszeichen: INST 1200/2-1 Version B1

Inhalt

Mit der Geschäftsstelle der DFG wurde abgestimmt, dass dieser Erfahrungsbericht zunächst als Materialsammlung und Basis für die noch erfolgende Publikation verfasst wird.

1. Ausgangssituation

In der SuUB Bremen wurde seit 2011 im Rahmen eines von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts die in Fraktur gedruckte Zeitschrift „Die Grenzboten“ (1841 – 1922) digitalisiert und durch konventionelle Fraktur-OCR-Verfahren im Volltext erschlossen. Die dabei erreichte OCR-Erkennungsrate von 98,6% sollte für die Anforderung der wissenschaftlichen Forschung verbessert werden, um nachfolgend als Volltext in das Korpus des Deutschen Textarchivs (Projekt Clarin D) der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) integriert zu werden. Im Rahmen des Projektes sollte die Qualität des OCR-Volltextes der Zeitschrift durch automatisierte Methoden zur Nachbearbeitung auf ca. 99,5% Textgenauigkeit verbessert werden, um damit eine Reduktion der Fehlerquote um 64% zu erreichen.

Das digitale Ausgangsmaterial als Ergebnis des Digitalisierungsprojektes der Zeitschrift „Die Grenzboten“ (Erscheinungsverlauf, Anzahl der Jahrgänge und Bände, Beschaffenheit des Schriftbildes usw.) wurde bereits ausführlich im Abschlussbericht in Abschnitt 2.3. ″Ausgangslage und Zielsetzung beider Projekte″ beschrieben.

2. Fazit

Im Rahmen dieses Projektes wurden zwei Systeme zur Nachkorrektur von OCR-Volltext entwickelt bzw. getestet. Zum einen wurde in der SuUB Bremen ein eigenes methodisches Verfahren „*Bremer Ansatz“* entwickelt, und zum anderen erfolgte in Zusammenarbeit mit der Firma ProjectComputing eine Adaption des bestehenden Webservices *OverProof*[1]*,* die auf deutschsprachigen Frakturtext angewendet werden kann.

Der im Rahmen des Projektes entwickelte *Bremer Ansatz* erwies sich als gut geeignet, die OCR-Volltextqualität des Textkörpers „Die Grenzboten“ deutlich zu ver-

[1] Das Angebot *OverProof* der australischen Firma ProjectComputing zur Korrektur von deutschem Frakturtext befindet sich derzeit noch im BetaStadium. http://overproof.projectcomputing.com/

bessern. Als Projektergebnis wurde prototypisch Software entwickelt, die grundsätzlich auch für die Qualitätsoptimierung anderer Frakturkorpora erfolgreich einsetzbar ist.

Die Zielvorstellung des Fortsetzungsprojektes, ausgehend von 98,6% eine Zeichenerkennungsrate auf bis zu 99,5% zu steigern, ließ sich nicht durchgehend erreichen. Es wurden schwankende Zeichenerkennungsraten im Intervall zwischen 98,72% und 99,50% erreicht, ausgehend von automatisiert berechneten Zeichenerkennungsraten, auf verschiedenen Abschnitten aus verschiedenen Jahrgängen mit verschiedenen Seitenanzahlen (vgl. Erfahrungsbericht Abschnitt 2.3.).

Die weiterhin im Projekt durchgeführte Analyse (AP 3,4) von automatisierten und teilautomatisierten Ansätzen zur Prozessierung von Volltexten beweist, dass die hier betrachteten Verfahren ein sehr effizientes Kosten/Nutzungsverhältnis bei der Optimierung der Volltextqualität - insbesondere bei größeren Digitalisierungsprojekten - aufweisen können.

Ein zentraler Erfolgsfaktor bei der halbautomatischen Korrektur von Frakturtexten liegt in der Verfügbarkeit entsprechender umfangreicher Lexika und historischer Wortformenlisten. Eine zukünftige gemeinsame Pflege und Weiterentwicklung solcher offener, frei verfügbarer Ressourcen wäre projektübergreifend sinnvoll. Die Erfahrungen im Rahmen dieses Projektes legen es nahe, zukünftig einen intensiven Austausch zwischen Korpora bereitstellenden Infrastrukturen und OCR-Nachkorrekturprojekten einzurichten. Auf diese Weise könnte die Qualität von Listen historischer Wortformen und damit auch von OCR-basierten Textkorpora kontinuierlich verbessert werden.

2.1. [Teil-]Automatisierung, Schnittstellen, Korrekturpotenzial und Aufwände

Die Schnittstellen von Digitalisierungsmanagementsystemen wie der *Visual Library* der Fa. Semantics bieten ausreichend Potenzial, um mit automatisierten Ansätzen OCR-Volltext nachzubearbeiten. Der automatisierte Zugriff auf Strukturdaten, auf Volltext und auf Images ist problemlos möglich. Es können auch einzelne Zeichen aus den Images pixelgenau ausgeschnitten werden. Der Umfang der auf diese Weise realisierbaren Korrekturen durch den *Bremer Ansatz* ist stark vom Umfang und der Qualität der dabei eingesetzten Liste der historischen Wortformen abhängig. Beide getesteten Korrektursysteme werden dabei durch Fehlermodelle parametrisiert, das System *OverProof* arbeitet zusätzlich mit einem Sprachmodell (siehe Erfahrungsbericht Abschnitt 2.3.).

Im Rahmen dieses Projektes wurden die jeweiligen Ansätze zur Ermittlung von Arbeits- und Zeitaufwänden zwar konzipiert und entwickelt, eine systematische Erfassung der verschiedenen Ansätze wurde jedoch nicht dokumentiert. Sollten die zwei im Projekt entwickelten Nachkorrektursysteme bei anderen Projekten zur Anwendung kommen, wären die Arbeits- und Zeitaufwände unter den jeweiligen Arbeitsbedingungen zu evaluieren.

2.2. Erfahrungen mit dem IMPACT Centre of Competence

Die Zusammenarbeit mit dem *IMPACT digitization.eu Centre of Competence*[2] (IMPACT CoC) hat diesem Projekt wesentliche Impulse gegeben. So wurde über das IMPACT CoC beispielsweise der Kontakt zur Firma *ProjectComputing* während der

[2] http://www.digitisation.eu/

DATeCH hergestellt. Im Anschluss ergab sich die in AP-6 beschriebene Zusammenarbeit mit John Evershed und Kent Fitch zur Anpassung des Webservices *OverProof* auf deutschsprachige Frakturschrift[3].

Im Rahmen der Evaluation war darüber hinaus der Einsatz des „Post Correction Tools“[4] aus dem EU IMPACT-Projekt vorgesehen. Nach genauerer Analyse beruhen Erfahrungen beim Einsatz dieses Tools jedoch bisher lediglich auf der Korrektur von wenigen hundert Seiten. Zudem ist bei der Nutzung dieses Systems ein erheblicher zusätzlicher manueller Aufwand notwendig, der bei der Korrektur des umfangreichen Textkorpus „Die Grenzboten“ mit über 185.000 Seiten nicht zur Verfügung stand.

2.3. Erfahrungen zur OCR-Nachkorrektur

Der *Bremer Ansatz* zur Nachkorrektur von OCR-Volltext (vgl. Abschnitt 4.7) basiert auf der Annahme, dass Textfehler überwiegend aus OCR-spezifischen Zeichenfehlern resultieren, wie z.B. Vertauschungen der Buchstaben u/n, e/c, s/f usw. Beim Abgleich gegen eine Liste historischer Wortformen (LdhW) war daher eine weitere Liste von Zeichensubstitutionen eine wesentliche Parametrisierung des Algorithmus. Weitere Parameter wurden aus den Frequenzen der Wortformen sowie aus den Frequenzen der Zeichenfehlertypen abgeleitet. Anders als beim Webservice OverProof ging bei dem *Bremer Ansatz* keine weitere Kontextinformation auf Wort- oder NGRAM-Ebene bei der eigentlichen Zeichenkorrektur ein.

Durch eine experimentelle Analyse auf der Basis eines aus dem Ground Truth-Volltext abgeleiteten hypothetisch vollständigen und fehlerfreien Wörterbuchs war es möglich, das grundsätzliche Potenzial des *Bremer Ansatzes* zu bestimmen. Danach ist mit dem *Bremer Ansatz* ausgehend von einer Zeichenerkennungsquote von 98,27% maximal eine Zeichenerkennungsquote von 99,22% erreichbar.

Beim *Bremer Ansatz* wurden nur Wortformen korrigiert, die selbst nicht in der Liste der historischen Wortformen enthalten waren. Damit ergeben sich ausgelassene Korrekturen (falsch-negative Korrekturen) wie „dem/dein“, „Hans/Haus“ etc., die nur über einen den Wortkontext berücksichtigenden Ansatz korrigiert werden könnten.

Zeichen- und Wortfehlerquoten für verschiedene Abschnitte aus verschiedenen Jahrgängen

		Ausgangsquoten		nach Korrektur	
Jahrgang	Seitenanzahl	Zeichen	Wörter	Zeichen	Wörter
1841 + 1842	352	98,27%	94,82%	98,72%	96,89%
1870	11	99,42%	98,26%	99,50%	98,19%
1900	9	97,52%	92,51%	98,76%	96,40%

Der algorithmische Ansatz des Webservice OverProof der Firma ProjectComputing enthält den Wortkontext berücksichtigende Korrekturkriterien[5]. Der Einsatz dieses Verfahrens lieferte zwar einen neuen Typ von falsch-positiven Korrekturen, insgesamt können die Korrekturergebnisse durch *OverProof* (vgl. Abschnitt 4.7) als mindestens gleichwertig bewertet werden. Es zeigte sich jedoch, dass die Analyse der Korrektur der *OverProof* Methode über die automatisiert ermittelten

[3] Siehe Details zu weiteren Erfahrungen mit dem IMPACT Centre of Competence in Abschnitt 4.8
[4] http://www.digitisation.eu/tools-resources/tools-for-text-digitisation/cis-lmu-post-correction-tool-pocoto/
[5] John Evershed and Kent Fitch. 2014. Correcting noisy OCR: context beats confusion. In Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14). ACM, New York, NY, USA, 45-51. DOI=10.1145/2595188.2595200 http://dx.doi.org/10.1145/2595188.2595200 (aufgerufen 27.04.15)

Zeichenerkennungsquoten wegen systematisch auftretender Korrekturfehler von einem Satzzeichen zu einem (bzw. mehreren) Leerzeichen zu weniger sicheren Ergebnissen führen.

Der von *OverProof* prozessierte Volltext enthielt zudem Zeilenenden, die nicht der Originalvorlage entsprachen. Daher wurde im Rahmen des Projektes für die Integration des nachkorrigierten Volltextes in das Deutsche Textarchiv (AP8) der qualtitativ bessere Volltext des Bremer Ansatzes verwendet.

2.3.1. Zeit- und Kostenaufwände

Ein Gesamtkorrekturdurchlauf aller 186.740 Dateien benötigte einen Zeitaufwand von ca. 4 Stunden und 15 Minuten auf einem aktuellen Desktop-PC[6]. Die Software des *Bremer Ansatzes* wurde für Parallelprozessierung entwickelt, d.h. alle verfügbaren Prozessoren werden beim Korrekturlauf gleichermaßen genutzt.

Die seitenweise angeforderte OCR-Nachkorrektur des gleichen Textmaterials durch den Webservice OverProof (gehostet auf einem Server in Deutschland) dauerte 3 Tage und 21,6 Stunden. Die Bearbeitung durch *OverProof* erfolgte im Rahmen des Projektes ohne Kostenberechnung und wurde als Teststellung betrachtet. Für die knapp 200.000 Seiten der Zeitschrift „Die Grenzboten“ hätte die OCR-Korrektur nach regulären Firmen-Konditionen mittels OverProof Kosten in Höhe eines niedrigen dreistelligen Betrages verursacht.

2.4. OCR-Nachkorrektur als Dienstleistung und Geschäftsprozess

Die Software zur OCR-Nachkorrektur (*Bremer Ansatz*) wurde zunächst für den Einsatz in diesem Projekt mit dem Ziel entwickelt, die Qualität des Volltextes der Zeitschrift „Die Grenzboten“ so zu verbessern, dass der wissenschaftlichen Forschung eine effiziente Übernahme in die Kollektionen des BBAW möglich wird. Gleichzeitig sollten die gesammelten Erfahrungen auch für andere ähnliche Projekte weitergegeben werden. Eine freie Veröffentlichung als Produkt bzw. Dienstleistung, die deutlich höhere Aufwände in der Softwareentwicklung und -pflege erfordert hätte, zählte nicht zum Bestandteil des Projektes.

Die Erfahrungen mit dem Webservice *OverProof* werden als vielversprechend zum Einsatz für andere Projekte beurteilt. Das marktfähige System, das als Webservice konzipiert ist, vermeidet Aufwände beim Auftraggeber in Bezug auf Servermanagement, Installation, Konfiguration und Systempflege. Als Standard für OCR-Volltext arbeitet *OverProof* mit Dateien im ALTO-Format. Im Rahmen der Kooperation wurde für dieses Projekt eine Anpassung für Dateien im ABBYY-XML-Format vorgenommen. Verfügbar ist derzeit ein Dienstleistungsangebot von *OverProof* mit Ablaufbeschreibung, Beispielen und Kostenrahmen für englischsprachige Nachkorrektur[7].

---

[6] 4 Prozessor (Intel® Core™ i5, 3 GHz) Desktop-PC mit 8 GB Hauptspeicher

[7] http://overproof.projectcomputing.com/about
Der Umfang der Dienstleistung wird als *Anzahl hochgeladener Wörter pro Monat* definiert. Dabei ist der Service für Massendigitalisierung ausgelegt. Das dokumentierte Kostenmodell reicht in den Bereich von einer Milliarde Wörtern, was ungefähr 2,3 Millionen Seiten entspricht. Die Kosten für den Seitenumfang der Zeitschrift „Die Grenzboten“ (ca. 187.000 Seiten) würde sich auf $5.420 beziffern. Zum Vergleich: Bei dem Erfahrungsaustausch mit anderen Bibliotheken und Projekten wurde in einer hypothetischen Beispielbetrachtung geäußert, dass die Projektkosten für eine Korrektur von 98,5% auf 99,5% Zeichenerkennungsquote für ein Projekt einer Größenordnung von 0,5 Millionen Seiten bei ca. 20.000 € liegen würden.

3. Entwicklungsperspektiven für die OCR-Nachkorrektur durch den *Bremer Ansatz* und Ausblick

Die Projekterfahrung hat gezeigt, dass sich eine weitere Optimierung von OCR-Nachkorrekturen nur dann erreichen lässt, wenn für verschiedene Jahrhunderte umfangreichere und qualitative hochwertige Listen von historischen Wortformen zur Verfügung stehen, die ohne Zugriffsbeschränkungen verwendet werden können.

Parametrisierung der Tokenisierung

Ein relativ häufiger Fehler im OCR-Korpus war das Vorkommen einer im deutschen Sprachraum des 19. Jahrhunderts ungebräuchlichen Variante eines Anführungszeichens „»“ (Beispiel "folge» / folgen"). Um diese Fälle korrigieren zu können, ist es notwendig, dass dieses Zeichen bei der Tokenisierung berücksichtigt wird und ein Teil des OCR-Vokabulars darstellt. Nur unter dieser Voraussetzung kann die Substitution "»/n" für den Korrekturalgorithmus aktiviert werden. Dieser Ansatz entspricht einer Parametrisierung der Tokenisierung, die beim Einsatz des *Bremer Ansatzes* für OCR-Korpora ermöglicht wurde.

Adressierung weiterer Typen von OCR-Fehlern

Der *Bremer Ansatz* kann bisher Fälle ausgelassener bzw. eingefügter Leerzeichen nicht korrigieren. Siehe hier einige Beispiele aus dem Grenzboten:

- erstauntenDeutschlaude
- Deutschlaudeinen
- Dentschlandsmüsse
- Dentschlandkann
- Dentschlandssorgen
- EntKicklungDentschlands;
- CentralisationDentfchlands

Der zusätzlich notwendige Aufwand zur Entwicklung eines dafür geeigneten Korrekturalgorithmus, der die Fälle verschmelzender und aufgeteilter Wörter berücksichtigt, wäre erheblich. Eine Abschätzung für das darin liegende weitere Korrekturpotenzial konnte im Rahmen des Projektes nicht ermittelt werden.

Weitere Desiderate

Eine weitere Optimierung des Bremer Ansatzes ließe sich erreichen, wenn verschiedene statistische Charakteristika in dem zu korrigierenden Textkorpus weiter untersucht würden. So könnte eine Analyse der Eigenschaften von hochfrequenten und niederfrequenten Wortformen (der sogenannte "*long tail*" der Wortfrequenzstatistik) zeigen, wie sich z.B. der *Bremer Ansatz* dagegen verhält. Angenommen wird, dass sich insbesondere zahlreiche Entitäten (z.B. Eigennamen, Fachbegriffe) und außergewöhnliche Komposita (wie z.B. Sonntagsröcklein oder Gewerbelehrlingsschulen) im „long tail“ befinden.

Auch die Identifikation von abweichenden Schriften (wie z.B. griechisch) und Sprachen bei der eigentlichen OCR oder auch bei der OCR-Nachkorrektur wäre wünschenswert.

Generell stellt sich die Frage, welches Korrekturpotenzial bei verschiedenen Niveaus von Zeichenfehlerquoten erreichbar ist. Die Ergebnisse zu verschiedenen Abschnitten aus dem „Grenzboten“ haben gezeigt, dass das Korrekturpotenzial bei höherer Zeichenfehlerquote steigt. So konnte zu einem Abschnitt aus dem Jahrgang 1900 die Zeichenfehlerquote von 2,48% halbiert werden. Hingegen konnte ein Abschnitt aus dem Jahrgang 1870, mit der geringsten Zeichenfehlerquote 0,58%, lediglich auf ein Niveau von 0,5% korrigiert werden.

Erweiterte Recherchemöglichkeiten in Textkorpora

Zur verbesserten Suche in digitalen Volltext-Korpora sollten differenzierte Recherchemöglichkeiten angeboten werden, dazu zählen insbesondere:

- Unscharfe Suche
- Suche mit Indizes vor und nach OCR-Nachkorrektur (eine Anregung der Firma ProjectComputing in Anlehnung an das Konzept „*union index*“ aus dem Projekt „*Australian Newspapers Online*“[8])
- Typische OCR-Fehler berücksichtigende Suche
- Suche mit Kontext-Wortwolke[9]
- Reguläre Ausdrücke

Kontextspezifische OCR-Prozessierung

Ein nicht quantifizierter, aber häufiger Fehler innerhalb des durch ABBYY berechneten Volltextes ist die Vertauschung von „u“ und „n“, der trotz eines offensichtlich vorhandenen Potenzials einer korrekten Erkennung im Schriftbild auftritt. So sind die Zeichen in der folgenden Abbildung Bestandteil des Ergebnisses einer Suche nach „Hans“[10].

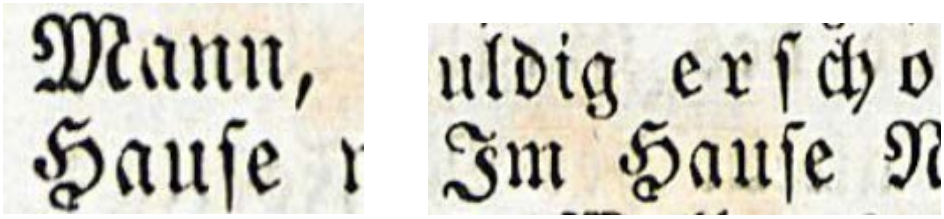
Es ist davon auszugehen, dass durch eine "kontextspezifische OCR-Prozessierung" für diese Fälle die Fehleranfälligkeit wie z.B. „Hans/Haus“ reduziert werden kann. Im Rahmen dieses Projektes konnte jedoch eine entsprechende Erweiterung des *Bremer Ansatzes* nicht mehr realisiert werden.

Noch als "u" erkennbare Zeichen wurden von der OCR als "n" in bestimmt

[8] http://trove.nla.gov.au/newspaper

[9] Unter einer Kontext-Wortwolke wird im Rahmen von Recherche-Szenarien eine über Synonyme hinausgehende Menge von semantisch „benachbarten“ Begriffen eines Kontextes verstanden. Die Relevanz in Bezug auf den Kontext wird über eine Gewichtung dargestellt.

[10] http://brema.suub.uni-bremen.de/grenzboten/periodical/pageview/279194?query=hans



Bildkontext für die Zeichen der vorgangegangenen Abbildung – VLID 279194

Zusammenfassend lässt sich feststellen:

- Der entwickelte *Bremer Ansatz* kann OCR-Volltexte gut korrigieren, so dass wesentliche Projektziele des Fortsetzungsprojektes mit Hilfe dieses Systems erreicht werden konnten.
- Bessere Lexika und historische Wortformenlisten würden das Ergebnis weiter optimieren.
- Bestimmte Fehler kann man nur mit einer (aufwendigen) kontextspezifischen OCR-Prozessierung beherrschen; so können beispielsweise ausgelassene Leerzeichen nur mit großem Aufwand korrigiert werden.
- Mit OverProof liegt ein vielversprechendes, kostengünstiges Serviceangebot im Betastadium vor, das zum Zeitpunkt des Projektantrages noch nicht bekannt war. Nach unseren Tests hat sich gezeigt, dass dieses System zukünftig deutsche Fraktur-OCR relativ einfach verbessern kann.