Korrektur von fehlerhaften OCR Ergebnissen durch automatisches Alignment mit Texten eines Korpus

Bald, Markus
markusbald92@gmail.com
Universität Würzburg, Deutschland

Damiani, Vincenzo
vincenzo.damiani@uni-wuerzburg.de
Universität Würzburg, Deutschland

Essler, Holger
holger.essler@uni-wuerzburg.de
Universität Würzburg, Deutschland

Eyeselein, Björn
bjoern.eyselein@uni-wuerzburg.de
Universität Würzburg, Deutschland

Reul, Christian
christian.reul@uni-wuerzburg.de
Universität Würzburg, Deutschland

Puppe, Frank
frank.puppe@uni-wuerzburg.de
Universität Würzburg, Deutschland

Einleitung

Bei der Transkription historischer Texte liefert die OCR (Optical Character Recognition) trotz deutlicher Verbesserungen mit Open-Source-Tools wie OCRopus 1.3 und Tesseract 3.5 bzw. 4.0 meist keine perfekten Ergebnisse. Oft ist der zu transkribierende Text jedoch schon an anderer Stelle verfügbar, ohne dass der genaue Fundort bekannt ist. Um die Nachkorrektur zu vereinfachen, muss zu dem zu transkribierenden Text der Vergleichstext gefunden und aligniert werden, um dann Abweichungen zu korrigieren (bei historischen Dokumenten ändert sich oft der Wortlaut eines Textes je nach Überlieferung in verschiedenen Quellen, so dass man nicht immer davon ausgehen kann, das jede Abweichung ein OCR-Fehler ist). Dafür stellen wir das Open-Source-Tool "OCR-Textkorpus-Aligner" vor. Es orientiert sich an der Vorgehensweise des Passim-Tools, geht aber darüber hinaus, indem u. a. der alignierte Text so aufbereitet wird, dass er auch zum Training der OCR-Software benutzt werden kann. Die Vorgehensweise besteht darin, zeilenweise mit N-Grammen (aktuell 5-Gramme) Kandidaten für ähnliche Zeilen im Vergleichstext zu generieren und die Zeile mit der höchsten Übereinstimmung als Ergebnis zurückzuliefern. Zusätzlich wird ein komfortabler Editor zur Nachkorrektur bereitgestellt. Der OCR-Textkorpus-Aligner wurde in zwei Teilprojekten im Kallimachos-Verbund-Projekt (www.kallimachos.de) erfolgreich eingesetzt: im Anagnosis-Projekt (http://www.kallimachos.de/kallimachos/index.php/Anagnosis:Main) und im Narragonien-Projekt (http://www.kallimachos.de/kallimachos/index.php/Narragonien:Main) für die Transkription von frühen griechischen Drucken und einer Druckausgabe des Narrenschiffes.

In Kap. 2 wird kurz der Stand der Forschung dargestellt und in Kap. 3 die Methoden und die Benutzungsoberfläche des Alignment-Tools "OCR-Textkorpus-Aligner" präsentiert. Kap. 4 beschreibt die Evaluationsergebnisse aus zwei Anwendungsdomänen, die in Kap. 5 diskutiert werden und Kap. 6 gibt einen Ausblick mit beabsichtigen Weiterentwicklungen.

Stand der Forschung

Das „Sequence Alignment“, eine musterbasierte Suche von ähnlichen Zeichenketten in großen Sequenzen, wird als wesentlicher Bestandteil der Bioinformatik hauptsächlich dazu verwendet, um ähnliche DNA-, RNA- und Proteinstränge zu finden, die auf strukturelle, funktionelle oder evolutionäre Beziehungen hindeuten können. Dementsprechend fokussieren sich die meisten bereits existierenden Alignment-Programme wie „Lalign“ (https://embnet.vital-it.ch/software/LALIGN_form.html) oder „BLAST“ (Altschul et al. 1997) auf die naturwissenschaftliche Anwendung, wobei sich durch die geringe Anzahl an Nukleotid-Buchstaben kaum Möglichkeiten zur Nutzung bei Textkorpora bieten. Eine Anwendung von BLAST für Textkorpora wird z. B. in (Vesanto et al. 2017) vorgestellt. Angepasste Versionen dieser Tools wie Passim (Smith et al. 2014) aus dem Leipziger "Open Philology Project", sind primär auf das Matching von längeren ähnlichen Textpassagen ausgerichtet, sodass häufig einzelne, insbesondere kurze Zeilen nicht gefunden werden können. Die Software verwendet als zentrale Komponente ein Framework namens „jAligner“ (Ahmed 2018), welches die am weitesten verbreiteten Sequence-Alignment-Algorithmen implementiert und dabei den zur Verfügung stehenden Zeichenvorrat nicht einschränkt.

Methoden

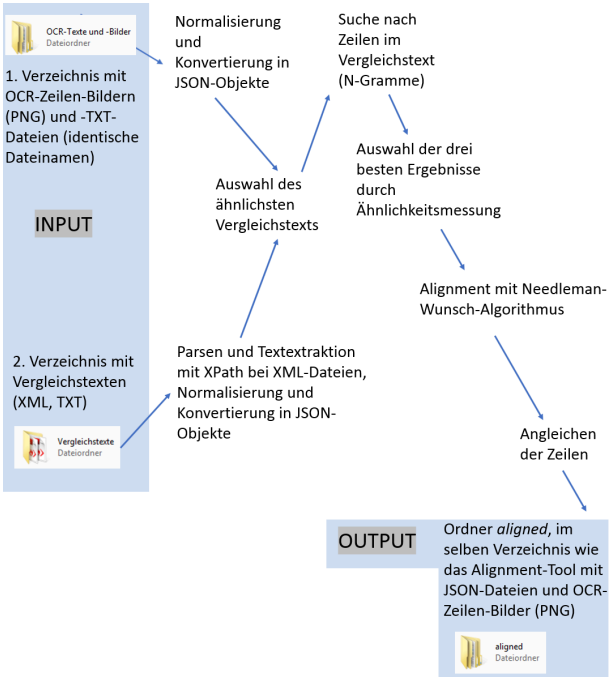

***Abbildung 1**: Workflow-Diagramm des Alignment-Tools vom Input (links) bis zum Output (rechts).*

Der Workflow des Tools "OCR-Textkorpus-Aligner" ist in Abbildung-1 veranschaulicht. Vor dem "Sequence Alignment" werden die Zeilen durch Entfernen der Diakritika normalisiert, um die Chance zu erhöhen, zu den OCR-Zeilen jeweils passende Entsprechungen im Vergleichstext zu finden. Durch eine Ähnlichkeitsmessung der Textanfänge wird zunächst das Vergleichsdokument mit der höchsten Entsprechung vorausgewählt. Anschließend wird jede zu transkribierende Zeile (im folgende "OCR-Zeile" genannt) in N-Gramme aus fünf Zeichen segmentiert und diese im Vergleichstext (im folgenden Ground-Truth bzw. GT-Zeile genannt) gesucht. Aus lokalen Clustern von Treffern bei der N-Gramm-Suche werden Kandidaten generiert, die hinsichtlich der Anzahl der gefundenen N-Gramme und einer Ähnlichkeitsmessung bewertet werden. Aus dieser Einschätzung ergibt sich der jeweils beste Kandidat. Der globale (über die volle Länge der Zeilen alignierende) Needleman-Wunsch-Algorithmus (Needleman und Wunsch 1970) richtet die OCR-Zeile und den besten GT-Kandidaten so aufeinander aus, dass möglichst viele Zeichen übereinstimmen. Fehlende Zeichen (z. B. ein Komma) in der OCR-Zeile im Vergleich zur GT-Zeile werden durch Trennstriche (-) aufgefüllt, die Lücken markieren. Dabei wird die Länge der OCR-Zeile an die Länge der GT-Zeile angepasst. Die Ergebnisse des Alignments dienen anschließend als Input für das Korrektur-Tool.

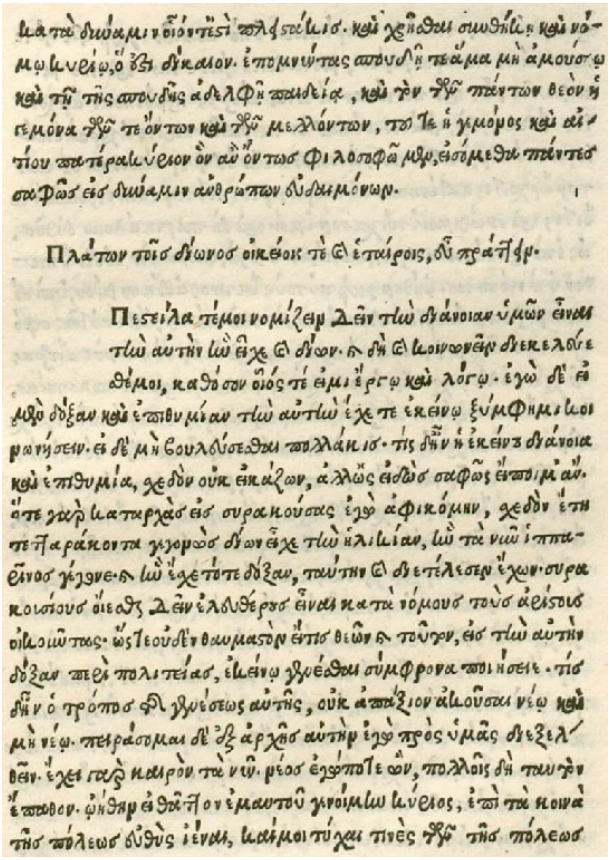

***Abbildung 2:** Eine Seite aus dem Band Epistulae diversorum philosophorum, oratorum, rhetorum, Venedig, A. Manutius, 1499 = GW 9367 [Platon, Briefe, 323c7-324c1].*

Ein Beispiel für eine Input-Seite findet sich in Abbildung-2. Abbildung-3 zeigt die Bedienung des OCR-Textkorpus-Aligner mit Alignment anhand der ersten beiden Zeilen von Abbildung-2. Unter der Anzeige des Orginal-Scans wird oben die GT-Zeile und darunter die OCR-Zeile angezeigt. Abweichungen einzelner Zeichen sind durch rote Kästchen markiert. Die Nutzer können dann entweder das Zeichen aus der GT-Zeile in die OCR-Zeile übernehmen oder umgekehrt oder mittels einer virtuellen, konfigurierbaren Tastatur beide Zeichen durch ein anderes ersetzen. Nach einer Markierung springt die Auswahl zur nächsten abweichenden Stelle weiter. Durch Tastenkombinationen lässt sich die Markierung der Fehler noch beschleunigen. Auch lassen sich alternative Alignment-Kandidaten auswählen und komplette Zeilen editieren. Die korrigierten Zeilen lassen sich herunterladen und als Ground Truth zum Training der OCR einsetzen.

Abbildung 3: *Oben: Startseite des OCR-Textkorpus-Aligner mit Auswahl des OCR-Ergebnisses und dem Ordner mit Vergleichsdateien. Mitte: Status des Alignment-Prozesses. Unten: Auswahl von den oberen beiden Zeilen aus Abbildung-2 zur Nachkorrektur in einem dreizeiligen Editor mit Originalzeile, durch Alignment ausgewählter Vergleichszeile ("GT") und OCR Zeile, wobei durch farbige Markierungen die Ground Truth festgelegt wird (ganz unten ein virtueller Editor mit domänenspezifischen Sonderzeichen).*

Evaluation

Bei den beiden Evaluationen wurde jeweils der Prozentsatz der korrekt alignierten OCR-transkribierten Zeilen zu allen Zeilen des OCR-Textes berechnet. Der erste Datensatz beinhaltet eine frühe Druckausgabe von 13 Briefen Platons in altgriechischer Sprache[1], verteilt auf zwölf Seiten. Die Scanbilder (Beispiel s. Abbildung-2) wurden in 302 Zeilen segmentiert, auf die anschließend eine OCR-Erkennung mit einer Fehlerrate von ca. 15% angewandt wurde. Als Vergleichstexte dienten 793 Transkriptionen der Perseus Digital Library. Diese ständig erweiterte Open-Source-Volltextdatenbank wird von der TUFTS University/Universität Leipzig auf Initiative von Prof. Gregory R. Crane seit 1987 entwickelt. Bei Nutzung der Software Passim lag bei unseren Experimenten die Erkennungsrate im besten Fall bei 136 von 302 Zeilen und damit bei 45% (mit N-Gramm-Größe auf Wortebene = 2 und Deaktivierung von einschränkenden Parameter wie eine Mindestanzahl an N-Gramm-Matches zwischen den Texten sowie eine Mindestlänge des Alignments). Demgegenüber hatte das „OCR-Textkorpus-Aligner“-Tool 292 von 302 Zeilen (ca. 96,7%) bei einer N-Gramm-Größe von fünf Zeichen korrekt zugeordnet. Eine Fehleranalyse der zehn nicht gefundenen Zeilen ergab, dass sich darunter vier zu kurze Zeilen (mit weniger oder knapp über fünf Zeichen) und sechs Zeilen mit sehr vielen OCR-Fehlern (wegen häufigen Großbuchstaben bzw. generell schlechter OCR-Qualität) befanden,

Der zweite Datensatz entstammt einer frühen Druckausgabe des mittelalterlichen „Narrenschiff“-Text (Ausgabe Basel vom 1.3.1497 = "GW 5061"), der im Rahmen des Würzburger „Narragonien“-Projekts digitalisiert wird, wobei die OCR-Fehlerrate ca. 18% betrug. Hier wurden von 10834 OCR-Zeilen 9384 GT-Zeilen im Vergleichstext einer anderen Druckausgabe des Narrenschiffs gefunden. Dies entspricht 86,62%, was angesichts von 1758 Zeilen mit kurzen Marginalien und 312 Zeilen, die lediglich aus Seiten-Nummern bestehen, ein gutes Ergebnis ist. Die Texte wurden vor dem Alignment durch Konvertierung in Kleinbuchstaben normalisiert.

Diskussion

Je weniger Zeichen eine Zeile enthält und je höher die OCR-Fehlerrate ist, desto schlechter ist das Alignment. Um diesen Zusammenhang zu quantifizieren, haben wir Erwartungswerte unter der idealisierten Annahme der Gleichverteilung der OCR-Fehler und konstanter Länge der Zeilen (10 bzw. 20 Zeichen) berechnet (s. Abbildung-4).

In den beiden Datensätzen der Evaluation enthielten die Zeilen meist mehr als 20 Zeichen, was bei einer OCR-Fehlerrate von 18% zu einem Erwartungswert von ca. 6% bzw. 33% mit keinem bzw. nur höchstens einem N-Gramm der Länge fünf zur korrekten Vergleichszeile führt. Bei einer OCR-Fehlerrate von 15% reduzieren sich diese Erwartungswerte auf 3% bzw. 23%. Wie erwartet beziehen sich die Fehler meist auf sehr kurze Zeilen, bei denen die Erwartungswerte deutlich höher sind (s. linke Kurve in Abbildung-4). Das Alignment lässt sich durch Normalisierung (z. B. bei Sonderzeichen oder Großbuchstaben) deutlich verbessern. Wir haben auch mit 4-Grammen statt 5-Grammen experimentiert, aber das ergab in den Beispieldomänen keine Verbesserungen, sondern nur höhere Laufzeiten, kann aber in anderen Domänen sinnvoll sein.

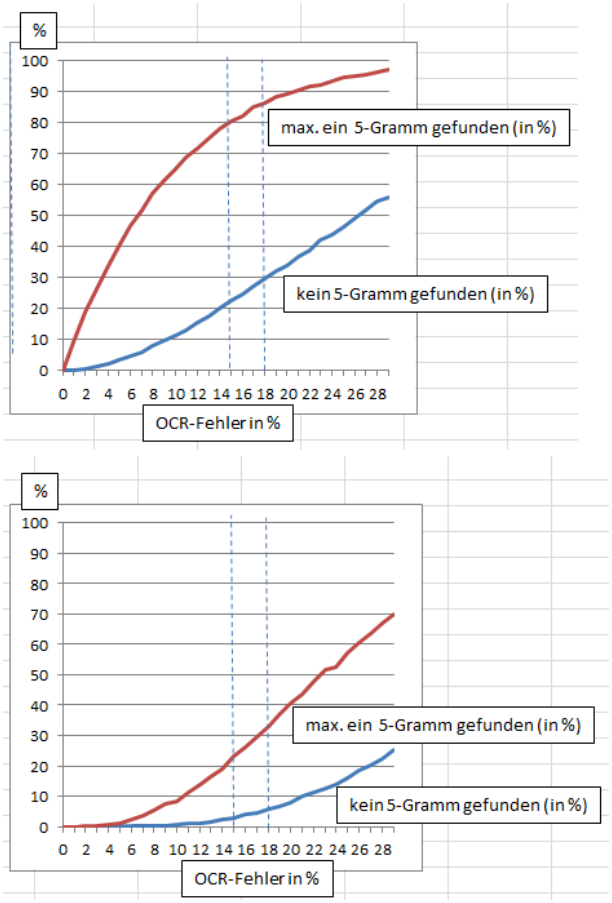

Abbildung 4: *Erwartungswerte für das Finden von keinem (untere blaue Linie) bzw. höchstens einem (obere rote Linie) 5-Gramm in einem Text mit 10 (linke Kurve) bzw. 20 (rechte Kurve) Zeichen, wenn die Fehlerquote der OCR-Erkennung für die 10 bzw. 20 Zeichen von 0% bis 29% variiert wird. Die gestrichelten Linien zeigen die Erwartungswerte bei einer Fehlerquote der OCR-Erkennung von 15% und 18% an.*

Zusammenfassung und Ausblick

Trotz relativ hoher OCR-Fehlerraten von 15% bzw. 18% konnten längere OCR-Zeilen relativ zuverlässig in GT-Vergleichstexten gefunden werden. Dafür reichen meist schon ein oder zwei passende N-Gramme aus. Um auch kürzere Zeilen zuordnen zu können, wollen wir die Tatsache ausnutzen, dass die Reihenfolge der Zeilen in OCR-Text und Vergleichstext meist übereinstimmt. Somit können aus der relativen Position von nicht erkannten Zeilen zu erkannten Zeilen Kandidaten für die korrekte Zuordnung generiert werden, die dann durch den Needle-Wunsch-Algorithmus auf Zeilenebene überprüft werden. Für den praktischen Gebrauch ist eine Einbindung in OCR Workflows wichtig. Dazu muss nur die Schnittstelle eingehalten werden, die auf Ordnern mit Dateien in Standard-Formaten sowie auf Namenskonventionen beruht (vgl. Abb. 1). Wir haben dies für das OCR-Framework „OCR4all“ umgesetzt, welches Algorithmen für alle Schritte von der Vorverarbeitung über die Segmentierung und die OCR einschließlich Nachtraining mit einem Editor zur Nachkorrektur bereitstellt.

Fußnoten

1. *Epistulae diversorum philosophorum, oratorum, rhetorum*, Venedig, A. Manutius, 1499 = GW 9367.

Bibliographie

Altschul, Stephen F. / Madden, Thomas L. / Alejandro A. Schäffer / Zhang, Jinghui / Zhang, Zheng / Miller, Webb / Lipman, David J. (1997): *“Gapped BLAST and PSI-BLAST: a new generation of protein database search programs”* in Nucleic Acids Res 25 (3389-3402).

Moustafa, Ahmed (2018): *“JAligner: Open source java implementation of Smith-Waterman”* in http://jaligner.sourceforge.net .

Needleman, Saul B. / Wunsch, Christian D. (1970): *“A general method applicable to the search for similarities in the amino acid sequence of two proteins”* in Journal of Molecular Biology 48, (443–453).

Smith, David A. / Cordell, Ryan / Dillon Maddock Elizabeth / Stramp, Nick / Wilkerson, John (2014): *“Detecting and modeling local text reuse”* in Proceedings of the JCDL’14 (Joint Conference on Digital Libraries; 183–192), IEEE Press.

Vesanto, Aleksi / Nivala, Asko / Salakoski, Tapio / Salmi, Hannu / Filip, Ginter (2017): *“A system for identifying and exploring text repetition in large historical document corpora”* in Proceedings of 21. NoDaLiDa (Nordic Conference on Computational Linguistics).