Das Mittelalter 2019; 24(1): 224–227

Tobias Hodel

# Lesende Algorithmen: Projekt READ

https://doi.org/10.1515/mial-2019-0016

Dank der Digitalisierung finden sich seit geraumer Zeit mehr und mehr digitale Repräsentationen mittelalterlicher Dokumente im weltweiten Netz. Die Arbeit der Mediävisten erfährt daher eine Verlagerung aus den Bibliotheken und Archiven vor die Bildschirme. Damit ändert sich als Konsequenz die Forschung mit Materialien, die in der Zeit vor der Erfindung des Buchdrucks entstanden. Deren Auffinden wird dank Beschreibungsdaten und Kataloginformationssystemen erleichtert. Digitale Editionen und digitalisierte Dokumente geben einen zusätzlichen Schub, unterschiedliche Texte zu vergleichen und zu vernetzen.

Eine weitere und tiefergreifende Umwälzung verspricht die Einführung der computergestützten Handschriftenerkennung. Diesem Unterfangen widmet sich das Projekt ‚Recognition and Enrichment of Archival Documents', kurz READ.[1] Unter dem Claim „READ Revolutionizes Access to Handwritten Documents" machen vierzehn europäische Partner handschriftliche Texte und frühe Drucke aus allen Zeiten und Regionen les- und durchsuchbar.

Mit Fokus auf die Mediävistik bedeutet der Einsatz von Automatisierungsprozessen, dass zukünftig Dokumente nicht nur aufgefunden, sondern Bilder und Texte, die digital vorliegen, durchsucht und ausgewertet werden können. Im Rahmen der aufgezeigten Vision sind dabei zwei Herangehensweisen zu unterscheiden, die gesondert betrachtet werden müssen, nicht zuletzt da sie unterschiedlich weit ausgereift sind: Einerseits die Suche in Dokumenten mit einer Volltextsuche ohne festen Volltext (sogenanntes *keyword spotting*), andererseits die zeichengetreue Transkription von Handschriften.

## Der transkribierende Computer

Die Vorgehensweise zur Erzeugung von automatisiert erstellten Transkriptionen umfasst zwei Stufen. Erstens muss ein Dokument segmentiert, also Textregionen

**1** Siehe die Projektwebsite: read.transkribus.eu. Alle Links wurden letztmals am 15.11.2018 abgerufen.

**Kontakt: Dr. des. Tobias Hodel,** Staatsarchiv des Kantons Zürich, Winterthurerstrasse 170, CH–8057 Zürich, Schweiz, E-Mail: tobias.hodel@ji.zh.ch

(bspw. Absätze) und einzelne Zeilen identifiziert werden. Die Segmentierung von handschriftlichem Material wurde in den vergangenen Monaten so stark verbessert, dass heute bei einfachem Layout nur noch sehr wenig manuelle Nachkorrekturen notwendig sind.[2] Als zweites muss ein Handschriftenmodell angewandt werden. Typischerweise trainieren Nutzende Handschriftenmodelle auf Basis von gleichen oder sehr ähnlichen Handschriften selbst. Das heißt, dem Algorithmus werden mehrere Seiten mit Transkription zur Verfügung gestellt, die dieser zur Optimierung von neuronalen Netzen braucht. Vermehrt stehen mittlerweile aber auch sogenannte generische Handschriftenmodelle zur Verfügung, die auf allgemeine Schrifttypen (bspw. gotische Buchminuskel) trainiert sind und auf entsprechende Schriften angewandt werden können. Für eine qualitativ hochwertige Erkennung bleibt indes die Bildung eines eigenen Modells vorrangig.

Erste Erfahrungen mit gotischen Buchhandschriften (13. und 14. Jahrhundert) sind ausgesprochen ermutigend. Bereits mit 40 Seiten Trainingsmaterial (ca. 35.000 Wörter) sind Zeichenfehlerquoten um 10 % erreichbar.[3] Die Zeichenfehlerquote (*Character Error Rate*, CER) weist die Anzahl an inkorrekt erkannten Zeichen nach, im Fall von zehn Prozent ist also jedes zehnte Zeichen, Leer- und Satzzeichen eingeschlossen, falsch. Noch vor zwei Jahren ist die Fehlerquote bei automatischer Erkennung doppelt so hoch gewesen.[4] Mit mehreren hundert Trainingsseiten lässt sich die Fehlerquote gar auf unter 3 % drücken.[5] Die Fehlerquote ist damit vergleichbar mit der automatischen Erkennung von Fraktur, die vor 1830 gedruckt wurde. Weitaus problematischer ist die Erkennung von Urkunden, da die Schriften stärker variieren und selten mehrere Dokumente von derselben Hand vorliegen.

Verstärkt wird aktuell an Modellen gearbeitet, die Trainingsdaten (also Bilder und Transkription) möglichst vieler unterschiedlicher Hände enthalten und dennoch brauchbare Fehlerquoten (Zeichenfehlerraten unter ca. 11 %) erzielen. Für Buchhandschriften ist dies bis zum Abschluss des Projekts sicherlich denkbar, für Urkunden und viele der Kursiven wird die Entwicklung noch etwas andauern.

**2** Tobias Grüning u.a., A Two-Stage Method for Text Line Detection in Historical Documents. In: Computing Research Repository 2 (2018). arXiv.org/abs/1802.03345v1.

**3** Siehe: Tobias Hodel, Medieval Handwriting and Handwritten Text Recognition. In: READ Blog 2017. https://read.transkribus.eu/2017/06/09/medieval-handwriting-and-handwritten-text-recognition/.

**4** Im Vgl. dazu: Eine Optical Character Recognition angewandt auf moderne Druckseiten wird ca. 0.5 % Zeichenfehler aufweisen.

**5** Der St. Galler ‚Parzival' (Cod. Sang. 857) lässt sich mit weniger als 1 % Zeichenfehlerquote erkennen, basierend auf einem Modell mit ca. 300 Seiten Trainingsmaterial.

Bereits heute lassen sich die Dokumente jedoch mit guten Erfolgsaussichten durchsuchen.

## Suche in Urkunden und Buchhandschriften

Für die Suche in unterschiedlichen Handschriften bietet sich der Einsatz des sogenannten *keyword spottings* an. Die Technologie nutzt nicht nur den erkannten Text, der typischerweise als beste Lesung (als erkannter Text) ausgegeben wird, sondern alle möglichen Zeichenvarianten, die durch ein Handschriftenmodell erkannt werden. Das heißt pro Sequenz (konkret pro Textzeile) wird eine Vielzahl von möglicherweise vorkommenden Zeichen erkannt und mit einer numerischen Konfidenz (Erkennsicherheit) versehen. Das Resultat ist eine Tabelle (eine sogenannte *confidence matrix*), die mit einer spezialisierten Suche, dem *keyword spotting* durchsucht werden kann.[6]

Auch mit Handschriftenmodellen, die nicht passgenau auf die zu erkennende Handschrift zugeschnitten sind und hohe Fehlerquoten in der Transkription erzeugen, lassen sich gute Suchresultate erzielen. Mit einem Modell, das eine – für die Transkription unbrauchbare – Fehlerquote von 25 % Zeichenfehlerquote generiert, werden noch mehr als 99 % aller möglicher Treffer gefunden (hoher *recall*). Die Anzahl an *false-positives*, also fälschlicherweise als Treffer angezeigte Resultate, steigt bei schlechteren Modellen jedoch an (niedrige *precision*).

Für westeuropäische Schriften aus dem Mittelalter (Runen ausgenommen) bestehen bereits heute zahlreiche Modelle, die für *keyword spotting* genutzt werden können. Entsprechend können große Dokumentenreihen mit der Methode durchsucht werden. Die Technologie lässt sich im Projekt ‚Himanis' anhand der handschriftlichen Register des französischen ‚Trésor des Chartes' austesten.[7]

Das Projekt READ wird durch die europäische Union im Rahmen des Forschungs- und Innovationsprogramm ‚Horizon 2020' gefördert und von der Abteilung Digitalisierung und Elektronische Archivierung der Universität Innsbruck koordiniert. Im Konsortium von READ sind unterschiedliche Forschungsrichtungen vertreten: Erstens, Forschungsgruppen aus dem Bereich der *Computer Vision*, die mehrheitlich mit der Layouterkennung befasst sind. Zweitens, Partner aus der angewandten Mathematik und der automatischen Sprachverarbeitung, die ihren Fokus auf Handschriftenerkennung legen. Drittens beteiligen sich schließlich

**6** Siehe dazu: Ioannis Pratikakis u. a., Keyword Spotting Engines: QbE, QbS. In: READ Deliverables 7/14 (2017). read.transkribus.eu/wp-content/uploads/2017/12/D7.14_v10.pdf

**7** Siehe die Projektwebsite: https://himanis.org/ und die Suchmaske: prhlt-kws.prhlt.upv.es/himanis/.

Geisteswissenschaftler, vornehmlich aus den Bereichen Geschichte und Archivwissenschaften an dem Unterfangen. Aus technischer Perspektive werden unterschiedliche Ansätze zur Erfüllung der Haupttasks (Layout- und Texterkennung) eingesetzt, das heißt die Aufgaben werden durch die verschiedenen Teams mit unterschiedlichen Lösungen bearbeitet, wobei intern ein ständiger Wettbewerb herrscht. Drei Archive, das Diözesanarchiv Passau, das Nationalarchiv von Finnland und das Staatsarchiv des Kantons Zürich, testen die entwickelten Algorithmen und die Infrastruktur mit eigenen Digitalisaten und beurteilen Qualität und andere Faktoren.

Die erfolgreichsten und massentauglichsten Lösungen werden in die Software und Infrastruktur ‚Transkribus' integriert, die in Innsbruck entwickelt wird. Bis zum Projektende Ende 2019 ist die kostenfreie Nutzung in jeglichem Umfang garantiert.[8] Zur langfristigen Verfügbarmachung wird danach eine europäische Kooperative für Forschungsprojekte, Institutionen und Einzelpersonen gegründet, die die Infrastruktur weiter tragen soll.

**8** Zu ‚Transkribus' siehe: transkribus.eu/Transkribus/.