Konsequenzen automatischer Texterkennung – Ein Aufriss zur Texterkennung mit Machine Learning

Hodel, Tobias
tobias.hodel@hist.uzh.ch
Staatsarchiv des Kantons Zürich, Schweiz

In den Geisteswissenschaften werden schon heute grosse Mengen an Text durchsucht, weiterverwertet und analysiert. Die Aufbereitung von Scans und Fotografien mit *optical character recognition* (OCR) bei gedruckten Texten wird erwartet. Gleichzeitig stehen selten Überlegungen im Zentrum, wie Text erkannt, welche Annahmen und vor allem welches Textverständnis vorausgesetzt wurde. Mit Hilfe des Sahleschen Textrads analysiert der Beitrag die automatisierte Erkennung von Texten mit neuronalen Netzen und fordert eine höhere Transparenz der verwendeten Methoden.

Sowohl Korpuslinguistik als auch Literatur- und Geschichtswissenschaften sind interessiert am Auffinden von Einzelbelegen, Mustern oder Entitäten in grossen Datenmengen. Tausende Seiten oder hunderte von Büchern lassen sich etwa mittels topic modeling klassifizieren (Schöch 2017). Der Prozess der Erkennung, der Weg zu den durchsuchbaren Texten, steht in den Überlegungen der Fächer jedoch meistens nicht im Zentrum. Obwohl Probleme der OCR-Erkennung angemerkt werden, ist die Textgüte nur bedingt ein Feld der Reflexion, die über Klagen hinausläuft (Ausnahme: Piotrowski 2012). Bedingt durch die Nutzung kommerzieller Produkte und entsprechend kaum offengelegter Prozesse, wird die Erkennleistung als gegeben angenommen und höchstens im *post-processing* die Qualität der Texte verbessert (bspw. PoCoTo: Vobl 2014).

Welcher Text?

Über das Monieren von Fehllesungen heraus, fehlt eine Reflexion, wie mit automatisch erkannten Texten umgegangen werden soll, gänzlich. Die Frage „welcher Text erkannt werden soll“, wird nicht thematisiert. Das hängt auch damit zusammen, dass Textverständnisse im digitalen Raum geprägt sind durch die Editionswissenschaften, einer Fachschaft, die aus einer anderen Richtung das digitale Feld bearbeitet. Die Qualität der Texterkennung, etwa der Transkription handschriftlicher Dokumente, steht nicht im Fokus, da bei menschlicher Erkennung durch Experten von einer Güte um 99,99% ausgegangen werden kann. Die Unsicherheiten, die unsicheren Lesungen, sind höchstens Teil von Paläographie orientierter fachspezifischer Debatten und nicht grundsätzlich ein qualitativer Messwert.

Die Editionswissenschaften waren es auch, die im Zuge der Digitalisierung Überlegungen zum Verständnis von Text hervorbrachten und sich – ganz im Sinne post-moderner Texttheorie – darauf einigten, dass es nicht den neutralen zu edierenden Text gibt. Ausgefaltet und visuell umgesetzt in Form eines „Textrads“ durch Patrick Sahle (Sahle 2013: 45-52). Erst das Verständnis von der Mehrschichtigkeit und Formbarkeit des Textbegriffes machte mehr oder minder konsequente Umsetzungen von digitalen Editionen überhaupt möglich und entspannte die Diskussion zwischen Philologie und Geschichtswissenschaft zu den je eigenen Vorstellungen von Text(-aufbereitung). Das Verständnis von Text unterscheidet sich dabei stark. Es kann vom Text als materiellem Ding ebenso ausgegangen werden, wie Text als Werk oder als Folge von Zeichen. Gerade für Fragen zur Umsetzung von Editionen, hilft das Textrad bei der Identifikation von Schwerpunkten, die Editionsentscheidungen unterstützen.

Im Rahmen von Projekten zur Texterkennung und Handschriftenerkennung durch grosse EU-Infrastrukturprojekte (IMPACT und READ) wurden derweil ebenso selten Überlegungen zum Text als Ressourcen angestellt und mehr auf Nachfragen beziehungsweise der Übernahme impliziter Vorstellungen abgestellt, die durch Informatiker oder Mathematiker bei der Entwicklung von Erkennungsalgorithmen eingebracht wurden. Dadurch muss auch deren Perspektive bei einer Theoretisierung der Texterkennung mitberücksichtigt werden.

Die Frage nach dem Verständnis von Text bei automatisierten Erkennvorgängen, hilft weiter bei der Abwägung zum Verhältnis von Mensch und Maschine beim Zugänglichmachen von Texten und stellt auch althergebrachte editorische Praktiken in Frage.

Automatisch erkannter Text

Die Anwendung des Sahleschen Textrads auf automatisierte Texterkennung macht deutlich, dass gewisse Textformen auch mit besten Erkennmethoden nicht isoliert werden können: Ausgangspunkt ist immer eine Textversion (Druck- oder Manuskriptseite), die in Form eines Faksimilie/Digitalisat vorliegen muss. Zeichenhaftigkeit und auch intellektuelle Bezüge sind daraus ableitbar, Text als Werk etwa, wie es rekonstruiert oder abstrahiert wird, lässt sich dagegen nicht erkennen.

Das digitalisierte Objekt agiert bei der automatisierten Erkennung jeweils als Ausgangspunkt, das erneut konsultiert werden kann und bei Kontrolle und Überprüfung hilft. Das bedingt jedoch, dass auch die Art und Weise der Digitalisierung (Auflösung und Farbechtheit, aber auch Format und Aufnahmeverfahren),

bei einer Kritik berücksichtigt werden müssen. Bereits der zugrundeliegende „Text“ ist also technisch geprägt.

Mit Fokus auf die Ausgabe des Erkennprozesses, werden erkannte Strings ins Zentrum gesetzt. Dabei muss nicht zwangsläufig nur ein String („die beste Lesung“) vorgelegt werden, sondern Varianten, also eine Reihe von Strings, die mehrere mögliche Lesungen enthalten, sind extrahierbar. Ergänzt um die durch die maschinell errechnete Wahrscheinlichkeit der Erkennung wird eine Matrix (sog. *confidence matrix*) an möglichen Lesungen und deren Wahrscheinlichkeit erstellt, die ebenfalls durchsucht werden kann. Insbesondere innerhalb von grossen Quellenmassen lassen sich so potente Suchen (Volltextsuchen ohne zugrunde liegendem Volltext sozusagen) realisieren, die ausgesprochen gute Ergebnisse erzielen. Mit dem Nachteil, dass je nach Suche auch *false-positive* Variantenlesungen vorgelegt werden. Die Methode wird daher nur bedingt für Auswertungen nutzbar, die auf Quantifizierung beruhen. Dank der Konfidenzen wird die Anzahl an Zeichenfehler, als zentralem Tool zur Messung von Textgüte relativiert, da ein alternativer Zugang zu den Strings besteht.

Innerhalb des Vorgangs zur Erkennung von Text kommt der eigentliche Erkennprozess jedoch erst an zweiter Stelle. Ebenso zentral und ebenfalls fehleranfällig, ist die Identifikation des Layout bzw. die Unterscheidung zwischen texttragenden und textfreien Zonen auf den zu erkennenden Digitalisaten, eine Aufgabe die für Menschen spielerisch einfach gelöst werden kann. Obwohl für reguläre Layouts von handschriftlichem Material in den vergangenen Monaten erhebliche Fortschritte erzielt wurden (Grüning 2017), bleiben Probleme im Umgang mit komplexen Layouts und insbesondere Tabellen, die in keine oder nur die ungenügende Identifikation von Zeilen mündet.

Der Punkt der Layouterkennung wird noch problematischer, da nur schwierig ausgewiesen werden kann, welche Teile als „texttragend“ identifiziert wurden. Bei Wettbewerben in den Computerwissenschaften werden unterschiedliche Messwerte angenommen. Etwa die Abweichung von einer manuell gezogenen Baseline [cBad] oder die Zuordnung zu Pixeln [DIVA-HisDB]. Allen Verfahren gemein ist der Bezug auf von Menschen hergestellte, relativ subjektive Grundlagen. Fakt ist, alles was nicht als Teil des Layouts identifiziert wird, kann im darauffolgenden Prozess nicht als Text erkannt werden.

(Vor-)Entscheidungen

Alle diese Überlegungen stellen nicht mehr als Grundlagen beziehungsweise Vorannahmen dar, die getroffen werden, bevor eine Erkennung überhaupt stattfinden kann. Im Gegensatz etwa zu händisch erstellten digitalen Editionen, ist eine Anpassung des Textbegriffs nicht im Erstellprozess möglich, sondern höchstens vor Beginn oder beim Abschluss der Bearbeitung.

Noch deutlicher wird die Abgeschlossenheit, nähert man sich der Texterkennung aus einer technischen Perspektive. Insbesondere für die Texterkennung von Handschriften und frühen Drucken lohnt sich der Einsatz von *machine learning*, konkret rekurrenten neuronalen Netzen (RNN). Das Training solcher Netze muss selbstredend vor der eigentlichen Erkennung erfolgen. Die trainierte Ausgabe entspricht dabei einem nachgeahmten, determinierten menschlichen Input. Je nach Grösse des Trainingssets und der Variabilität der Schriften wird dies mehr oder minder genau erreicht. Zentral im Prozess ist das erwartete Resultat, oder anders formuliert die Art und Weise, wie Text aufbereitet wird. Die Aufbereitung selbst, beispielsweise die stillschweigende Auflösung von Abkürzungen, Normalisierung von Schreibungen oder das Einfügen bzw. Zusammenführen von Zeilenumbrüchen, wird Konsequenzen auf die Ergebnisse haben. Auch der verwendete Zeichensatz (etwa die Codierung in Unicode) oder Vereinheitlichungen wird das Resultat beeinflussen.

Aus technischer Sicht gibt es innerhalb des Trainingsprozesses selbst nur einige wenige Parameter, die kontrolliert werden können. Entsprechend ist der gesamte Rest, Teil einer *blackbox*, die auch nicht näher analysiert werden kann, da das Funktionieren der einzelnen Neuronen in einem Netz nur schwer und häufig ohne Einsichten zum Funktionieren der gesamten Erkennung beobachtet werden können.

Ein Kontrollmechanismus findet sich einzig im standardisierten Testen der trainierten Modelle mit Hilfe von Testsets, also nach gleichem Muster hergestellte Seiten, die nicht fürs Training verwendet wurde und entsprechend Auskunft über die Leistungsfähigkeit eines Modells geben können. Zentrale Messwerte dabei sind *Character Error Rate* und *Word Error Rate*, die auf so genannter Ground-Truth basiert, also von Menschen hergestellte „korrekte“ Lesungen der Texte.

Eine weitere Form der Einflussnahme besteht in der Verwendung von Wörterbüchern, die bei Unsicherheit herangezogen werden und plausiblere Lesungen (= im Wörterbuch) gegenüber anderen Strings bevorzugen. Für historische Schreibformen bestehen zwar Korpora, jedoch ist der Einsatz für Texte vor Ende des 19. Jahrhunderts (insbesondere für vormoderne Texte) umstritten, da keine Konventionen bestanden und die Gefahr der Hyperkorrektur aufgrund des verwendeten Wörterbuchs besteht.

Auch wenn es bislang nur beschränkte Erfahrungen mit dem Einsatz von *machine learning* bei der automatischen Erkennung von Text gibt, ist absehbar, dass die Verbesserungen zum Einsatz der Technologie führen werden. Mit dem Preis, dass Probleme des *machine learnings* mit eingeführt werden. Die *biases*, insbesondere die Perspektiven im Moment der Aufbereitung, von Trainingsdaten etwa, werden übernommen (Zundert 2016: 341). Im Kontext von automatisierter (Vor-)Aufbereitung von Bewerbungsdossiers oder nicht gender-gerechten Auswertungen von Informationsmassen wird das Probleme des Datenbias bei Methoden des *machine learnings*

rasch sichtbar (Siehe dazu einen jüngeren Artikel aus dem britischen Guardian, Devlin 2017). Bei der Texterkennung mögen die Konsequenzen gesellschaftlich weniger gravierend ausfallen, problematisch und kritisch zu analysieren sind sie nichtsdestotrotz.

Ansprüche an automatische Texterkennung

Wie der kurze Abstecher in die Welt des maschinellen Lernens zeigte, ist eine Kontrolle der Erkennleistung nur ganz bedingt und basierend auf wenigen Faktoren möglich, entsprechend lässt sich zusammenfassend im Umgang mit automatisierten Texterkennungsalgorithmen eine Reihe von Forderungen ableiten, damit erkannte Texte kritisch eingeordnet werden können.

Messwerte basierend auf Testsets müssen ausgewiesen werden: Character Error Rate und Word Error Rate geben Aufschlüsse zur Qualität des erkannten Textes. Darüber hinaus ist eine Einschätzung sinnvoll, in welchem Bereich Falschlesungen häufig identifiziert wurden (Eigennamen, Zahlen etc.). Insgesamt sollte dadurch einsichtig werden, in welchen Bereichen Qualitätsprobleme zu erwarten sind.

Standards und Kernfragen an die publizierten Texte offen dokumentieren: In den Editionswissenschaften bereits praktiziert, wird der Umgang mit Frageperspektiven verbesserte Einsichten liefern, vor welchem Hintergrund Textkorpora erstellt wurden. Daran schliesst sich die Forderung nach **Offenlegung der zugrunde liegenden Ground-Truth zur Erstellung von Test- und Trainingssets** an: Damit wird nachvollziehbar, was zur Modellerstellung genutzt und auch, welchen Standards, Richtlinien und Gepflogenheiten dabei gefolgt wurde.

Durch den kritischen Umgang mit automatisch erkannten Texten eröffnet sich ein fundierter Umgang mit denselben, der mit gewissen Sicherheiten eine Weiternutzung von Text ermöglicht und die textzentrierten Teile der *digital humanities* in eine kritikfähige Zukunft führt.

Bibliographie

Devlin, Hannah (2017): AI programs exhibit racial and gender biases, research reveals. In: The Guardian vom 13.04. URL: https://www.theguardian.com/technology/2017/apr/13/ai-programs-exhibit-racist-and-sexist-biases-research-reveals.

Grüning, Tobias, Labahn, Roger, Diem, Markus, Kleber, Florian, Fiel, Stefan (2017): READ-BAD: A New Dataset and Evaluation Scheme for Baseline Detection in Archival Documents. [Preprint submitted to DAS2018] arXiv:1705.03311. URL: https://arxiv.org/abs/1705.03311.

Piotrowski, Michael (2012): Natural language processing for historical texts. Morgan & Claypool, San Rafael.

Sahle, Patrick (2013): Digitale Editionsformen. Zum Umgang mit der Überlieferung unter den Bedingungen des Medienwandels. Teil 3: Textbegriffe und Recodierung., Schriften des IDE. BoD, Norderstedt.

Schöch, Christoph (2017): Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. Digital Humanities Quarterly 11.

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. ACM, New York, NY, USA, pp. 57–61. doi:10.1145/2595188.2595197

Zundert, Joris J. van (2016): Screwmeneutics and Hermenumericals: The Computationality of Hermeneutics, in: Schreibman, Susan, Siemens, Ray, Unsworth, John (Eds.), A New Companion to Digital Humanities. John Wiley & Sons, pp. 331–347.