Pre-circulated Paper für die Tagung *digital history*: Konzepte, Methoden und Kritiken digitaler Geschichtswissenschaften, Göttingen/online 1.-3.3.2021

von Tobias Hodel, Walter Benjamin Kolleg, Digital Humanities, Universität Bern

Die Maschine und die Geschichtswissenschaft:
Der Einfluss von *deep learning* auf eine Disziplin

Es gibt wenige Technologien, die die Phantasie der Menschen so stark beflügeln, wie sie gleichzeitig durch ihren Einsatz beeindrucken können. *Deep learning* gehört zweifelsohne dazu. Das Verfahren aus dem Bereich des maschinellen Lernens wird mittlerweile für jegliche Bewertungsentscheide eingesetzt, die vor wenigen Jahren noch als ungeeignet für die Bearbeitung durch Algorithmen oder allgemein „den Computer" beurteilt wurden.

Das Prinzip des *deep learning* ist einfach, aber nicht unproblematisch: Neuronale Netze, dem menschlichen Gehirn nachempfundene vernetzte Speicherzellen, werden mit möglichst vielen vorgefertigten Daten versorgt und in einem Trainingsprozess auf die zu lösende Aufgabe getrimmt. Entscheidend sind zwei Größen: Die Quantität und die Qualität der eingegebenen Daten. Von Spracherkennung, über Bildanalyse zu Dokumentenauswertung – *deep learning* setzt sich als Technologie in diversen Feldern durch und wird seit wenigen Jahren auch für naturwissenschaftliche Auswertungen benutzt.

In den Geisteswissenschaften wird die Technologie aktuell erst in Ansätzen genutzt. Die Texterkennung von Drucken und Handschriften ist nur einer von vielen Einsatzbereichen, der sich diese Ansätze zu Nutze macht. Bereits absehbar ist indes, dass in naher Zukunft weit mehr (Be-)Wertungsentscheide manuell unterstützt oder gar autonom getroffen werden. Named Entity Recognition, aber auch visuelle und textuelle Strukturanalysen zeigen gemäß ersten Tests und *proof-of-concepts* bessere Resultate, als dies rein regelgeleitete Algorithmen vermögen. Mit wenig Phantasie lassen sich gar die Einsatzmöglichkeiten noch erweitern und die Interpretation von Texten mit und dank *machine learning* modellieren.

Im Rahmen dieses Papers werden drei Themenblöcke angeschnitten, die unterschiedliche Anwendungen des maschinellen Lernens im Fokus haben. Erstens, und wohl am unproblematischsten, ist die Nutzung von *deep learning* zur Handschriftenerkennung. Der menschliche *bias* fließt zwar durch Transkriptionsentscheide in die Automatisierung ein, führt aber zu harmlosen Fehlern und Hyperkorrekturen. Problematischer ist Zweitens, die Entitätenerkennung (*Named Entity Recognition*), die kulturwissenschaftliche Fragen zu Praktiken der Namensgebung und zum Individuum im generellen aufwirft. Drittens können schließlich mit *machine learning* Ansätzen Strukturerkennung betrieben werden – dies ist eine Vorgehensweise, die in analoger Form etwa aus der Urkundenlehre bereits bekannt ist. Ein entsprechendes Training mit Übernahme der kanonisierten Wertung ist möglich, führt aber unweigerlich zur Verstärkung impliziter und expliziter Bevorzugungen. Im deutschsprachigen Raum wären solche Bevorzugungen etwa die Prägung der Diplomatik durch die Analyse ausgefertigter Königsurkunden im Gegensatz zu den zahlenmäßig massiv überwiegenden Urkunden, die unter dem Label „Privaturkunden" zusammengefasst werden. Um die Technologie in den Fokus zu stellen, ist es jedoch nötig, dass die drei Ansätze innerhalb des Arbeitens mit neuronalen Netzen verortet und vor allem die Resultate kritisch betrachtet werden. Dieses Paper orientiert sich daher an den drei Perspektiven «Training» von neuronalen Netzen, «Interpretation» von Input und Output, sowie «Konsequenzen» des Einsatzes.

Trainieren: Die Induktion von *bias*

Training als Basis zur komplexen, statistisch unterstützten Wertung erweist sich als größte Stärke und gleichzeitig neuralgische Stelle der Aufbereitung, da durch das Trainingsmaterial (Vor-)Urteile übernommen und verstärkt werden. Diese Effekte wurden etwa für Suchmaschinen oder bei Bewerbungsprozessen mehrfach nachgewiesen und problematisiert.[1] Je nach Form des maschinellen

[1] Noble, Safiya Umoja: Algorithms of Oppression: How Search Engines Reinforce Racism, New York 2018; Caliskan, Aylin; Bryson, Joanna J.; Narayanan, Arvind: Semantics derived automatically from language corpora contain human-like biases, in: Science 356 (6334), 14.04.2017, S. 183–186. Online: <https://doi.org/10.1126/science.aal4230>.

Lernens werden Trainingsmaterialien vorgegeben und überprüft (*supervised learning*) oder Strukturen selbständig erlernt (*unsupervised learning*).

Im Rahmen der **Handschriftenerkennung** werden Bildausschnitte einem zu erkennenden Text gegenübergestellt. Die Aufgabe des neuronalen Netzes ist es eine Entsprechung zwischen Anhäufungen an Pixeln und zu erkennenden Zeichen zu finden. Dabei agieren die meisten Systeme unabhängig vom Vorwissen und trainieren jeweils eigenständige Modelle. Damit gibt es keine «natürliche» Verbindung zwischen Zeichen und Bild, es hängt vielmehr von den Vorgaben ab, die im Rahmen des Trainings gemacht werden. Die Distinktion von Zeichen liegt bei der trainierenden Person und mehr noch im Umgang mit einem Zeichen in den vorhandenen Trainingsdaten.

Analog dazu verhält es sich bei der **Erkennung von benannten Entitäten**, einem Problem, das ebenfalls mit *supervised* Ansätzen bearbeitet wird und einen Algorithmus zur Nachahmung verleiten soll. Das System versucht *tokens* («Wörter» im Satzkontext) als einer Gruppe (Person, Ort, Organisation, etc.) zugehörig zu bestimmen. Im Vergleich zur Erkennung von handschriftlichem Text basiert ein zentraler Schritt auf der Anwendung von Sprachmodellen. Solche Modelle können große Textmengen in mehrdimensionalen Vektorräumen verorten und damit Ähnlichkeiten zwischen Wörtern aufzeigen, weil diese entweder häufig im selben Kontext auftauchen oder aus ähnlichen Teilen bestehen. Beim Training der Entitätenerkennung wird einem System entsprechend vermittelt, inwiefern Wörter im Umfeld eines Vektors zu einer gemeinsamen Gruppe gehören. Wiederum sehen wir, wie die Aufbereitung von Trainingsmaterial zur Induzierung von (Vor-)Urteilen führt, indem etwas als Person oder Ort verstanden wird. Als zusätzliches Problem stehen wir vor der Herausforderung, dass Sprache einem ständigen Wandel unterworfen ist, der die Erzeugung von historischen oder domänenspezifischen Sprachmodellen erforderlich macht. Diese Modelle sind natürlich aufgrund ihrer Basis auch «gefärbt» (sog. «Korpusfärbung») und bilden Sprache nur entsprechend dem zugrundeliegenden Korpus ab.

Wenn wir noch einen Schritt weiter gehen und als drittes versuchen **Textteile einem Thema** zuzuweisen oder nach semantischen Gesichtspunkten zu segmentieren, bewegen wir uns sowohl im Bereich des *supervised* als auch des *unsupervised learning*. Neben dem Trainieren von Annotationen (etwa für *sentiment analysis*) können auch Sätze (selten Satzteile) aufgrund der vorkommenden Wörter zu Themenfeldern zusammengefügt werden (etwa mit *topic modeling*, Blei, 2011). Text wird dabei jeweils als «bag of words» verstanden, die Reihenfolge der Wörter also ignoriert. Als zentrale Einheit gilt in diesen Verfahren der Satz, der durch einen Punkt von der nächsten Einheit abgetrennt wird. Entsprechend lässt sich an dieser Stelle ein neuralgischer Punkt identifizieren, da viele vormoderne Sprachen keine Entsprechung zum Satz kennen. Das zweite Problem ist die Aufbereitung der zu identifizierenden Teile oder Themen, da durch die Identifikation von Urkundenteilen (Protokoll, Kontext oder Eschatokoll) diese je nach Ausgangsmaterial unterschiedlich stark gewichtet werden.

Die drei kurz skizzierten Themenbereiche stellen unterschiedliche Phasen im Prozess der Quellenaufbereitung dar. Dabei zeigt sich sowohl in relativ simplen Erkenn- oder Identifikationsprozessen als auch in komplexen Zuordnungen das Moment des Trainings als kritischer Vorgang, da die daraus generierten Modelle je nach Korpus (Ausgangsmaterial) in einen Modus des Nachahmens übergehen. Das Verständnis der Modelle ist in der Konsequenz ein hermeneutischer Prozess, der wie bereits von Gadamer gefordert (Gadamer, 2010) eine Auseinandersetzung mit (eigenen) Urteilen und insbesondere Vorurteilen miteinschließt und folglich den Prozess des Trainings nur so nachvollziehbar macht.

Interpretieren: Quellenkritik und Hermeneutik

Über den Prozess des Trainings hinaus, stellt der Umgang mit Resultaten des *machine learning* insbesondere aus Sicht der *algorithm studies* eine Herausforderung dar. Die Algorithmen lassen sich zwar an unterschiedlichen Stellen zu Ausgaben zwingen (bekannt sind die Google Image-Traum Algorithmen, Mordvintsev et al., 2015), jedoch ist ein Nachvollzug der Entscheide innerhalb neuronaler Netze bislang nicht erfolgreich möglich. Die Kritik und die Auswertung der Resultate aus Vorgängen des maschinellen Lernens ähneln entsprechend hermeneutischen Interpretationen, die gerade durch den geschichtswissenschaftlichen Werkzeugapparat wie der Quellenkritik, aber auch anderen Methoden

analysiert werden müssen.[2] Erst das wechselseitige *close-* und *distant-reading* der Quellen und der Resultate macht es möglich, die Belastbarkeit der maschinell gewonnenen Wertungen zu überprüfen.

Die Überprüfung der Fehler ist bei der Texterkennung auf den ersten Blick relativ simple, da relativ standardisierte Transkriptionskonventionen existieren. Neue Erkennalgorithmen erreichen dabei, je nach Anzahl der Trainingsseiten, unterschiedliche Resultate. Für Handschriften ist eine Erkennqualität mit Fehlerquoten im Bereich von 2,5% technisch möglich, pro 1'000 erkannten Zeichen wird entsprechend mit 25 Fehlern gerechnet, worunter auch die fehlerhafte Erkennung von Satz- sowie Groß-/Kleinschreibung fällt. Bei regelmäßigen Schriften lässt sich diese Fehlerquote durch das Training eines entsprechenden Modells mit ungefähr 50'000 Wörtern erreichen.[3] Das Resultat lässt sich unter optimalen Bedingungen, d.h. genügend Material von ähnlichen Schriften, auch für Modelle erreichen, die auf unterschiedlichen Händen basieren.

Es bleibt die Frage offen, inwiefern durch die Quantifizierung von Fehlern Aussagen zur Leistungsfähigkeit eines Erkennmodells gemacht werden können. Zentral bleibt aus historischer Perspektive schließlich die Fragestellung und die (digitale) Methode, die nach dem Erkennprozess zum Einsatz kommen soll. Je nachdem fällt auch der Fehlertyp (Satzzeichen sind für *topic modeling* Algorithmen etwa unerheblich) oder die Art eines Fehlers (die Verwechslung von Stab-s mit „f" führt im *close reading* zu keiner/wenig Verwirrung) ins Gewicht. Zukünftig wird es entsprechend wichtig über quantifizierende Fehlerquoten hinaus, Angaben zur Fehleranfälligkeit eines Modells zu machen.

Der Einsatz von *named entity recognition* verlangt anders gelagerte Diskussionen. Wie bereits oben angesprochen, wird auch dabei der Trainingsinput imitiert. Dies basiert auf Sprachmodellen, sodass auch die kritische Analyse eines solchen Modells Teil der Methodenkritik wird. Bei der Anwendung historischer Sprachformen besteht zusätzlich das Problem, dass Sprachmodelle auf verhältnismäßig kleinen Datenmengen basieren.

Um die Leistungsfähigkeit bestehender Frameworks für nicht-standardisierte vormoderne Sprachen zu demonstrieren, wurde im Rahmen des Editionsprojekts Königsfelden ein Experiment zur Erkennung benannter Entitäten durchgeführt. Dabei wurde ein eigenes Sprachmodell (selbsttrainiert als FLAIR embeddings)[4] angelegt, das auf zeitlich nahen Dokumenten basiert.[5] Das Training der benannten Entitäten basiert auf 645 Urkunden, für Verhältnisse des maschinellen Lernens also insgesamt eher wenig Material. Eine Besonderheit bildet das Tagging des Editionsprojekts, das die Strategie verfolgt, alle potentiell zugehörigen Informationen einem Namen zuzurechnen. Dadurch wurden auch Angaben, die heute nicht mehr als Namensteil verstanden würden, als solcher markiert und folglich auch fürs Training verwendet. Trotz der wenig Trainingsdaten konnten F-Scores im Bereich von 69-74% erreicht werden.[6] Auch für dieses Verfahren lohnt sich ein Blick auf einzelne Resultate. Somit lässt sich nämlich eine Vielzahl von „Fehlern" sichtbar machen, die korrekte Resultate widerspiegeln. Die „Fehler" stammen in dem Fall von Annotator*innen, die inkorrekt auszeichneten oder aber die Maschine liefert gar valable alternative Annotationen (Namen können teilweise Orts- oder Personennamen bezeichnen). Einschränkend muss erwähnt werden, dass die Transkription händisch erstellt und Eigennamen im Gegensatz zum restlichen Text großgeschrieben wurde. Die Algorithmen hatten entsprechend starke Indizien zur Identifikation von Entitäten.

Stärker noch als bei der Texterkennung zeigt sich für Annotationsaufgaben, wie sehr die unterschiedlichen Inputs (Sprachmodell, Transkriptionsvorgaben und Trainingsmaterial) das Resultat beeinflussen. Eine Analyse der Technologie und der Resultate muss die Komplexität mitberücksichtigen, wobei aktuell der benötigte «Werkzeugkasten» dazu noch mehrheitlich fehlt und wiederum quantitative Angaben nur beschränkt Aussagen zur Fähigkeit eines Netzes erlauben.

---

[2] Für die Literaturwissenschaften siehe als Beispiel: Underwood, Ted: Emerging conversations between literary history and sociology., in: The Stone and the Shell, 02.12.2015, <https://tedunderwood.com/2015/12/02/emerging-conversations-between-literary-history-and-sociology/>, Stand: 11.04.2019.

[3] Für dieses und andere Beispiele siehe (Hodel, 2020)

[4] FLAIR ist ein open source Framework für Natural Language Processing: https://github.com/flairNLP/flair.

[5] Verwendet wurde das Bonner Frühneuhochdeutsch Korpus (http://www.korpora.org/FnhdC/), digital vorliegende Bände der Schweizerischen Rechtsquellen (https://www.ssrq-sds-fds.ch/home/) und Urkunden und Akten des Klosters Königsfelden (https://www.hist.uzh.ch/de/fachbereiche/mittelalter/lehrstuehle/teuscher/forschung/projekte/koenigsfelden.html).

[6] F-Scores kombinieren Recall (Ausbeute) und Precision (Präzision) und sind ein häufig genutztes Mittel, um Klassifikatoren zu beurteilen.

In einem weit experimentelleren Stadium als die Identifikation von Entitäten befindet sich die Zuordnung von Annotationen, die Sinneinheiten klassifizieren. Bereits etwas etabliert, vor allem da kommerziell interessant, ist die *sentiment analysis*, die Sätzen positive und negative Gefühlsausdrücke zuordnet. Analog dazu können auch andere, etwa thematische Labels vergeben und trainiert werden. Die bereits oben beschriebenen Probleme werden dabei übernommen und die Komplexität nochmals um eine Stufe gesteigert, da die Sprachmodelle auf der Ebene «Satz» angewandt werden.

Da gleichzeitig typischerweise auf „den Satz" als einfach zu segmentierende Einheit Bezug genommen wird, entstehen etwa für vormoderne Texte oder wenig gepflegte textuelle Formen (Stichwort Kurznachrichten) Herausforderungen. Gerade für die bereits angesprochenen Urkunden ist der Satz keine sinnvolle Einheit, um Zuordnungen zu erstellen.

Mit Blick auf diesen dritten Themenbereich (Textteile einem Thema zuordnen), stehen wir heute in einer initialen Findungsphase. Erste Modelle führen zu vielversprechenden Eindrücken jedoch nur wenig belastbaren Resultaten. Auch die Anwendung von *topic modeling* auf einzelne Sätze ist möglich, führt aber zu einem Clustering von ähnlichen Wortkonstruktionen und mahnt an die Auswertung von Kookkurrenzen. Aufschlüsse zu semantischen oder gar thematischen Feldern werden damit nur mittelbar gegeben.

Konsequenzen: Von einer neuen Heuristik zu einer neuen Epistemologie?
Das Oszillieren zwischen praktischen Umsetzungen und theoretischen Überlegungen, führt zu neuen Problemstellungen, die Epistemologie und heuristische Methoden der Geschichtswissenschaften betreffen. Maschinelles Lernen zeigt sich dabei bereits heute als nützliche Erweiterung der Disziplin an der Schwelle des Einsatzes von *big data*, die es indes kritisch zu betrachten und zu verfolgen gilt. Die Einsichten dienen dabei nicht nur der intradisziplinären Methodendiskussion, sondern führen darüber hinaus zu kritischen Positionen für den Einsatz von *deep learning* im alltäglichen Leben.

Die Nutzung von *deep learning* in einer hochgradig reflexiven Wissenschaft wie der Geschichtswissenschaft, bedeutet die Explizierung erkenntnistheoretischer Grundannahmen. Was etwa als «Text» verstanden wird, muss offengelegt sein, wenn ein Algorithmus zur Erkennung von «Text» gebracht wird. Dabei regen auch einzelne Vorstufen die Diskussion an, wenn etwa identifiziert werden muss, wo sich auf einem Artefakt Text befindet. Die Identifikation von Personen oder Orten in textuellen Strukturen bindet ebenso an Vorannahmen zurück, indem Fragen nach bedeutungstragenden Namen gegenüber von Zuschreibungen abgewogen werden müssen. Zentral wird die Dokumentation der Aufbereitung der Grundlage, die Bewertungsentscheide nachvollziehbar macht. Indirekt lassen sich darauf aufbauend die Entscheide eines Modells nachvollziehen.

Die Anwendung von maschinellen Lernverfahren erfordert somit nicht eine komplett neue historische Methode, sondern eine Erweiterung des technischen Horizonts, indem zumindest im Grundsatz die Verfahren verstanden werden müssen. Überdies ist eine konsequente Erweiterung der Hermeneutik auf eingesetzte Methoden notwendig, da nicht mehr nur das erforschte Material, sondern auch die technischen Herangehensweisen nie vollständig überblickt und auch nur in (langsamer) Annäherung verstanden werden können.

Der *machine learning turn* führt nicht zu einer Abkehr von der historischen Methode, sondern vielmehr zu einer neuen Art der Beschäftigung mit Quellen, die nicht nur den Aussagewert beurteilt, sondern gleichzeitig auch die (automatisierte) Beschäftigung damit berücksichtigt.

Bibliografie

Blei, D. M. (2011). Introduction to probabilistic topic models. *Communication of the ACM*.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Gadamer, H.-G. (2010). *Hermeneutik I: Wahrheit und Methode: Grundzüge einer philosophischen Hermeneutik* (7th ed., Vol. 1, p. XII, 495). Mohr Siebeck.

Hodel, T. (2020). Best-practices zur Erkennung alter Drucke und Handschriften – Die Nutzung von Transkribus large- und small-scale. In C. Schöch (Ed.), *DHd 2020. Spielräume Digital Humanities zwischen Modellierung und Interpretation* (pp. 84–87). https://doi.org/10.5281/zenodo.3666689

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press.

Underwood, T. (2015, December 2). Emerging conversations between literary history and sociology. *The Stone and the Shell*. https://tedunderwood.com/2015/12/02/emerging-conversations-between-literary-history-and-sociology/