

Deep Learning als Herausforderung für die digitale Literaturwissenschaft

Zusammenfassung

In den Digital Humanities (DH) wird seit einigen Jahren über einen „computational turn“ (Berry 2011, Heyer 2014) diskutiert, der die aktuellen Algorithmen und Computertechniken des maschinellen Lernens und des tiefen Lernens stärker berücksichtigt. Beispiele sind das auf word embeddings basierende *word2vec* Modell, das darauf trainiert ist, sprachliche Zusammenhänge von Wörtern zu rekonstruieren (Mikolov 2013), oder *fastText* und *GloVe*, die ebenfalls zur Textklassifikation auf der Grundlage von word embeddings erstellt wurden. Diese Technologien haben sich in verschiedensten Bereichen der Verarbeitung von Bilddateien, Online-Datenbanken, der Erkennung natürlicher Sprache, prosodischer Daten und verschlüsselter Textdaten bewährt. Dennoch fehlt es in der Literaturwissenschaft und insbesondere der Textanalyse (Prosa, Lyrik, Drama) bis heute an einer Anwendung dieser neuen Methoden des maschinellen bzw. tiefen Lernens.

Bisher konzentrierten sich die klassischen „Digital Humanities“ eher auf die Generierung und Reflexion digitaler Ressourcen wie Textausgaben, Repositorien oder Bilddatenbanken. Dagegen gibt es nur wenige Versuche, Deep Learning in die digitalen Geisteswissenschaften einzubringen. Zumeist wurde Deep Learning in sehr großen Datenbanken von Unternehmen wie Google, YouTube, Bluewin Labors oder Echonest getestet, etwa um Social Media Signale und den Inhalt von Medien in sozialen Netzwerken zu analysieren. Gerade deshalb blieb in diesem Feld die alte Kluft zwischen traditionellen Geisteswissenschaften und Informatik bestehen. Unser Panel will einen Beitrag leisten, um diese Lücke zu schließen.

Wir wollen vor allem die Probleme erörtern, die bei der rechnerischen Analyse literarischer Texte mit Techniken des tiefen Lernens entstehen, z.B.: Können maschinelle Lerntechniken durch Clustering tatsächlich verdeckte Muster in Textdaten erfassen (Graves 2012)? Wie lassen sich auf der Grundlage eines maschinell erlernten Modells Grenzfälle, Kategorisierungsfehler, Ausreißer und ähnliche Besonderheiten erkennen bzw. in den Klassifikationsprozess einbauen? Wie geht man mit dem großen Problem der „black box“ um, wie lassen sich die in den „hidden layers“ stattfindenden Klassifikationsprozesse nachvollziehen bzw. gar transparent machen? Und welche Tools für die manuelle (z.B. Sonic Visualizer) und

automatische Annotation (z.B. PRAAT, ToBI, oder Sphinx) bzw. welche Softwares für die Modellierung (DyNet, TensorFlow, Caffe, MxNet, Keras, ConvNetJS, Gensim, Theano, und Torch) sind empfehlenswert?

Fragestellung und Aufbau des Panels

Wir haben auf unserem Panel Experten für computergestützte Analysen von literarischen Texten (Prosa, Lyrik, Drama) mit Interesse an vertieften Lerntechniken versammelt, die die computationale Analyse anhand von narrativen (Fokalisierung), dramatischen (Aktantenanalyse), poetischen (Metrik, Metaphorik, Reim) oder gattungsübergreifenden (Stilometrie, Topic Modeling) Textmerkmalen bereits umfangreich erprobt haben. Wir wollen uns vor diesem Hintergrund über bewährte Verfahren, praktische Anwendungen und erlernbare Methoden des Deep Learning austauschen, eine Plattform zur Präsentation und Entdeckung laufender Forschungsprojekte bieten, und die Vorteile und potenziellen Mängel der digitalen Mustererkennung auf der Grundlage der Methoden des tiefen Lernens reflektieren.

Folgende Personen haben dabei eine Teilnahme an dem Panel zugesagt:

- Fotis Jannidis ist Professor für Computerphilologie und Neuere Deutsche Literaturgeschichte an der Julius-Maximilians-Universität Würzburg.
- Christof Schöch ist Professor für Digital Humanities an der Universität Trier und Co-Direktor des Trier Center for Digital Humanities.
- Jonas Kuhn ist Professor für Maschinelle Sprachverarbeitung an der Universität Stuttgart und leitet das Centrum für Reflektierte Textanalyse (CRETA)
- Thomas Haider ist wissenschaftlicher Mitarbeiter am Max-Planck-Institut für empirische Ästhetik in Frankfurt am Main.
- Timo Baumann ist Informatiker am Language Technology Institute der Carnegie Mellon University in Pittsburgh, USA.
- Hussein Hussein ist Informatiker an der Freien Universität Berlin.
- Burkhard Meyer-Sickendiek ist Leiter einer von der Volkswagenstiftung geförderten Forschergruppe im Bereich der maschinellen Prosodieerkennung von Hörgedichten.

Wir werden in einem ersten Teil von maximal 30 Minuten einzelne Impulsvorträge präsentieren, und dann in einem zweiten Teil von ebenfalls 30 Minuten Topic Modeling und Embedding als wichtige Themenfelder des tiefen Lernens in den Geisteswissenschaften fokussieren. In einem dritten Teil von ebenfalls 30 Minuten wollen wir dann das Panel für die Diskussionen mit dem Publikum öffnen.

Panel-Vorträge

1. Thomas Haider vom Max Planck Institut Frankfurt wird über Topic Modeling am Beispiel der Latent Dirichlet Allocation (LDA) sprechen. Sein praktisches Beispiel ist die Anwendung der LDA auf einen Korpus neuhochdeutscher Poesie (Textgrid, mit 51k Gedichten, 5M-Token), um auffällige Themen in ihrem zeitlichen Verlauf (1575-1925 n. Chr.) und deren Verteilung ausfindig zu machen, und diese wiederum hinsichtlich der Klassifizierung von Gedichten in Epochen und mögliche Autorschaften auszuwerten. Die meisten Themen sind leicht interpretierbar und zeigen einen klaren historischen Trend. Um dennoch robustere Ergebnisse zu erhalten, plädiert Haider für eine verbesserte Methodik zur Berechnung der Wahrscheinlichkeit eines Themas bei einem bestimmten Zeitfenster. Für ein exploratives Experiment ist die Klassifizierung in Zeitfenster und Autorschaft zwar vielversprechend, für bessere Themenmodelle im Sinne klarer Trend- und Top-Themen wäre jedoch eine bessere Grundlage für die Modellierung diachroner Transformationen in der Poesie notwendig.
2. Fotis Jannidis von der Universität Würzburg wird dann darüber sprechen, wie man Techniken des Word Embeddings zur Modellierung literarischer Texte verwenden kann. Word Embeddings ermöglichen eine Repräsentation von Worten, die diese in eine gut interpretierbare Relation zueinander setzen: räumliche Nähe kann als semantische Ähnlichkeit verstanden werden. Mit Word Embeddings lassen sich zudem historische Entwicklungen von Worten anhand der historischen Veränderungen ihrer nächsten Nachbarn darstellen. Diese semantischen Netzwerke für die Ideengeschichte lassen sich auch im Rahmen literarischer Gattungen wie Prosa, Drama und Lyrik, etwa zum Vergleich verschiedener Themenfelder im Bereich der Lyrik nutzen. Interessant sind auch die Möglichkeiten zur Textrepräsentation: Während gängige bag-of-words-Modelle Texte über gewichtete Wortfrequenzen repräsentieren, gehen diese Ansätze von n-dimensionalen Embeddings aus und erreichen eine Textrepräsentation durch Summierung oder Durchschnittsbildung u.a.m. Durch die rasante Entwicklung der Embedding-Verfahren – insbesondere im Jahr 2018 – ergeben sich hier ganz neue Möglichkeiten für die digitalen Geisteswissenschaften.
3. Anschließend wird Christof Schöch von jüngeren Anwendungen von Deep Learning in den Literaturwissenschaften berichten, die von verschiedenen Arbeitsgruppen durchgeführt wurden. Diese Anwendungen verwenden teils Word Embeddings für die interne Repräsentation der Texte, zeichnen sich darüber hinaus aber vor allem durch den Einsatz künstlicher neuronaler Netze mit einer vergleichsweise hohen Anzahl von Schichten aus (vgl. Goodfellow et al. 2016). Diese Architekturen sind für einen großen Teil der Merkmalsmodellierung und für die Lösung der jeweils in Frage stehende Klassifikationsaufgabe verantwortlich. Dabei reichen die Anwendungskontexte von Optical Character Recognition für historische Drucke (u.a. Wick et al. 2018) über linguistische Annotationen (vgl. Kestemont und De Gussem 2017) und die Erkennung von Figurenrede in Romanen (vgl. Jannidis et al. 2018) bis hin zur Autorschaftsattribuion (vgl. Kestemont et al. 2016) und der Generierung von Gedichten (Hopkins und Kiela 2017). Während die Performanz der auf Deep Learning basierenden Ansätze häufig beeindruckend ist, werfen sie zugleich Fragen der Interpretierbarkeit und damit auch der Relevanz für die fachwissenschaftliche Theoriebildung auf.
4. Jonas Kuhn von der Universität Stuttgart wird das Verhältnis zwischen Deep Learning und theoriegeleiteter Forschung thematisieren. Im Zentrum stehen jüngere Erfolge mit Deep-Learning-Ansätzen, die auf der Grundlage eines sogenannten End-to-end-Training auf großen Mengen von realen Eingabe-/Ausgabepaaren (beispielsweise Maschinelle Übersetzung oder automatische Spracherkennung) basieren. In Hinblick auf eine Optimierung der Vorhersagequalität für diese Aufgaben werden klassische theoretische (Zwischen-)Konstrukte obsolet (für die Übersetzung etwa eine symbolische Repräsentation der syntaktischen Struktur oder der wörtlichen bzw. einer kontextuell implizierten Bedeutung). Aber werden theoretische Analysemodelle (typisch umgesetzt etwa im maschinellen Lernen auf handannotierten Trainingsdaten) damit insgesamt obsolet? Dies hängt davon ab, ob man eine Systemoptimierung für spezifische Aufgaben, oder aber ein wissenschaftliches Erkenntnisinteresse verfolgt. Gleichwohl stellt sich gerade bei Forschungsfragen der Digital Humanities, die oftmals unterschiedliche etablierte Beschreibungsebenen überspannen, die Frage der Motivation für etwaige symbolische Zielkategorien ganz neu.
5. Burkhard Meyer-Sickendiek und Hussein Hussein von der Freien Universität Berlin sowie Timo Baumann von der CMU Pittsburgh werden über character embeddings in der Gedichtanalyse sprechen. Kann man prosodische bzw. rhythmische Muster in moderner Lyrik mittels neuronaler Netztechniken auf der Grundlage von embeddings identifizieren? Für solche Klassifikationen werden feature-basierte Verfahren und Techniken tiefen Lernens vergleichend untersucht (vgl. Escudero et al. 2017). Dem merkmalsbasierten Ansatz, der auf Lyriktheorien im Bereich der freien Versprosodie basiert, wird dabei ein neuronaler Ansatz gegenübergestellt, anschließend wird die Integration von höherem Wissen in das neuronale Netzwerk erläutert. Unser Projekt arbeitet mit character embeddings auf der Grundlage bidirektionaler,

rekurrenter neuronaler Netzwerke (RNN, unter Verwendung von Gated Recurrent Unit (GRU)-Zellen (Cho et al. 2014)), die die Sequenz von Zeichen in eine mehrdimensionale Darstellung überführen. Es trainiert also nicht auf Wortebene, kann jedoch Informationen (z.B. Wortteile) über die konstituierenden Zeichenfolgen dekodieren. Dabei werden wir über Techniken des End-to-End-Lernens (Hannun et al., 2014; Graves and Jaitly, 2014) sprechen, die z.B. in der Spracherkennung verwendet werden, die weder Phoneme noch Wörter explizit modelliert, sondern direkt Audio-Features auf Charakter-Streams überträgt.

Diskussion

Mögliche Themengebiete für die Paneldiskussion im dritten und letzten Teil wären insbesondere die Verwendung tiefer Lernverfahren in den digitalen Geisteswissenschaften, etwa mit Blick auf Stilometrie, Computerstilistik, Reim- und Metrikenanalyse, Aktantenanalyse, oder Themenmodellierung. Dabei soll die Publikumsdiskussion all jenen ein Forum bieten, die sich ein tieferes Verständnis und eine praktische Schulung in deep learning sowie eine Plattform für den Austausch von Praktiken, Ergebnissen und Erfahrungen im Umfeld mit einschlägigen Tools erhoffen. Dies kann sich auch auf Kenntnisse aus den Nachbardisziplinen erstrecken, insofern diese über bereits vorhandenes Wissen hinsichtlich der Anwendung „tiefer Lerntechniken“ etwa im Bereich des Data Mining, der Statistik oder der Verarbeitung natürlicher Sprache verfügen. Auf diese Weise erhoffen wir uns eine effektive Fokusverlagerung innerhalb der digitalen Geisteswissenschaften: von der Erstellung und Archivierung digitaler Artefakte und Repositorien hin zu echten Rechenlösungen auf der Grundlage maschinellen Lernens.

Bibliographie

Ananthakrishnan, Sankaranarayanan / Narayanan, Shrikanth S. (2008): “Automatic prosodic event detection using acoustic, lexical, and syntactic evidence” IEEE Transactions on Audio, Speech, and Language Processing, 16(1), pp. 216-228.

Berry, David M. (2011): “The Computing Turn: Thinking About the Digital Humanities”, in: Culture Machine. 12, 1-22.

Brett, Megan R. (2012): “Topic Modeling: A Basic Introduction. Journal of Digital Humanities”.

Cho, Kyunghyun / Merrienboer, Bart van / Gulcehre, Caglar / Bahdanau, Dzmitry / Bougares, Fethi / Schwenk, Holger / Bengio, Yoshua (2014): “Learning phrase representations using RNN encoder-decoder for statistical machine translation”. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language.

Eder, Maciej / Kestemont, Mike / Rybicki, Jan (2013): “Stylometry with R: a suite of tools”. In: Digital Humanities, Conference Abstracts, University of Nebraska, Lincoln, 487-89.

Escudero, David / González, César / Gutiérrez, Yurena / Roder, Emma (2017): “Identifying characteristic prosodic patterns through the analysis of the information of Sp_ToBI label sequences”. Computer Speech & Language 45: 39-57.

Goodfellow, Ian / Bengio, Yoshua / Courville, Aaron (2016): “Deep Learning”. Cambridge, Massachusetts: The MIT Press.

Graves, Alex (2012): “Supervised sequence labelling with recurrent neural networks” (Vol. 385). Springer.

Graves, Alex / Jaitly, Navdeep (2014): “Towards end-to-end speech recognition with recurrent neural networks”. In International Conference on Machine Learning, pages 1764-1772,.

Hannun, Awni / Case, Carl / Casper, Jared / Catanzaro, Bryan / Diamos, Greg / Elsen, Erich / Prenger, Ryan / Satheesh, Sanjeev / Sengupta, Shubho / Coates, Adam (2014): “Deep speech: Scaling up end-to-end speech recognition”. arXiv preprint arXiv:1412.5567.

Hasegawa-Johnson, Mark / Chen, Ken / Cole, Jennifer / Borys, Sarah / Kim, Sung-Suk / Cohen, Aaron / Zhang, Tong / Choi, Jeung-Yoon / Kim, Heejin / Yoon, Taejin (2005): “Simultaneous recognition of words and prosody in the Boston University Radio Speech Corpus”. Speech Commun 46: 418-439, 2005.

Heyer, Gerhard (2014): “Digital and computational humanities”, in: http://dhd-wp.hab.de/files/book_of_abstracts.pdf, pp.66f.

Hopkins, Jack / Douwe Kiela (2017): “Automatically Generating Rhythmic Verse with Neural Networks”. ACL. <http://dx.doi.org/10.18653/v1/P17-1016>

Hsu, Chih-Wei / Lin, Chih-Jen (2002): “A comparison of methods for multi-class support vector machines”. IEEE Transactions on Neural Networks, 13:415-425.

Jannidis, Fotis / Konle, Leonard / Zehe, Albin / Hotho, Andreas (2018): “Analysing Direct Speech in German Novels”. DHd Jahrestagung. <http://dhd2018.uni-koeln.de/wp-content/uploads/boa-DHd2018-web-ISBN.pdf>

Jeon, J. / Liu, Y. (2009): “Semi-supervised learning for automatic prosodic event detection using co-training algorithm”. In Proc. of the 47th Annual Meeting of the ACL, pp. 540-548.

Kestemont, Mike / De Gussem, Jeroen (2017): “Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning”. Journal of Data Mining & Digital Humanities. <http://arxiv.org/abs/1603.01597>.

Mikolov, Tomas / Chen, Kai / Corrado, Greg / Dean, Jeffrey (2013): “Efficient estimation of word representations in vector space”. arXiv preprint. arXiv:1301.3781.

Moretti, Franco: Distant reading. Verso 2013.

Schöch, Christof: Quantitative Semantik: Word Embedding Models für literaturwissenschaftliche

Fragestellungen. In Digitale Literaturwissenschaft, hrsg. Fotis Jannidis. Stuttgart: Metzler (im Druck).

Wick, Christoph / Reul, Christian / Puppe, Frank: Calamari (2018): “A *High-Performance Tensorflow-Based Deep Learning Package for Optical Character Recognition*.” ArXiv:1807.02004 [Cs] <http://arxiv.org/abs/1807.02004> , 2018.