Spielräume bei der retroperspektivischen Analyse der Wittgenstein-Edition und die Herausforderungen für das Semantic Clustering

Hadersbeck, Maximilian
maximilian@cis.uni-muenchen.de
Ludwig-Maximilians Universität München

Ullrich, Sabine
sabine.ullrich@campus.lmu.de
Ludwig-Maximilians Universität München

Still, Sebastian
sebastian.still@campus.lmu.de
Ludwig-Maximilians Universität München

Pichler, Alois
Alois.Pichler@uib.no
Wittgenstein Archiv Universität Bergen/Norwegen

Einleitung

Seit 2010 kooperieren das Wittgenstein Archiv an der Universität Bergen (WAB, Alois Pichler) und das Centrum für Informations- und Sprachverarbeitung der Ludwig-Maximilians Universität München (CIS, Max Hadersbeck et. al.) in der Forschungsgruppe „Wittgenstein Advanced Search Tools" (WAST). Die WAST-Projektgruppe entwickelt die web-basierte FinderApp WiTTFind (http://wittfind.cis.lmu.de/), die einen computerlinguistisch gestützten digitalen Zugang zu WABs Wittgenstein-Edition erlaubt. Nach einer kompletten Neuscannung des Nachlasses und intensiven Verhandlungen des WAB mit den Rechteinhabern, dürfen seit 2018 WABs Edition auf der WiTTFind-Webseite durchsucht und Faksimileextrakte dargestellt werden. Nun konnten wir uns einer zentralen Frage der Wittgensteinforscher widmen: Wo finden sich in seinem Nachlass semantisch ähnliche Bemerkungen und, retroperspektivisch betrachtet, wann fanden diese Änderungen statt?

Wir entwickelten das Analysetool WiTTSim (Ullrich, 2018), das semantisch ähnliche Bemerkungen in der Edition aufspürt, zusammen mit einem vorgeschalteten semantischem Clusterverfahren (Ullrich, 2019), welches die Rechenzeit der Ähnlichkeitssuche um den Faktor 100 verkürzte. Zur retroperspektivischen Analyse der Edition entwickelten wir ein zeitorientiertes, textgenetisches Datenmodell, das die Spielräume der Interpretation der bisher dokumentorientierten Edition auf zugelassene Lesarten reduziert.

In unserem Vortrag stellen wir die Verfahren unserer Ähnlichkeitssuche mit vorgeschaltetem semantischen Clustering und ein neues mehr textgenetisch- als dokumentorientiertes Modell einer Edition vor, das im Web-Frontend des OdysseeReaders (www.odysseereader.wittfind.cis.lmu.de) implementiert ist und auch die Frage beantwortet: „Wann gibt es semantisch ähnliche Bemerkungen".

Die Datenbasis: Dokument- und Zeitorientierte Modelle

Die bei uns verwendete Datenbasis BNE 2015- und IDP 2016-, die am Wittgensteinarchiv an der Universität Bergen (Pichler, WAB) erstellt werden, enthalten Faksimile und Transkriptionen (auf der Basis von XML-TEI-P5) des Nachlasses von Ludwig Wittgenstein. Dieser Nachlass umfasst ca. 20.000 Seiten, welche vom WAB in Dokumente und diese wiederum in logische Textabschnitte unterteilt sind. Jeder der 54.930 Textabschnitte – eine sogenannte Bemerkung – wird mit einer eindeutigen Bezeichnung, dem sogenannten Siglum, versehen und wird in unserer Ähnlichkeitssuche als einzelnes Textobjekt definiert und semantisch analysiert.

Betrachtet man die Annotationen der BNE unter dem Aspekt der Retroperspektive, taucht folgendes Problem auf: Die BNE liefert nur auf der Ebene der Bemerkungen Informationen über ihren Erstellungszeitpunkt bzw. -zeitrahmen. Die Änderungen auf Wort und Zeichenebene sind zwar akribisch annotiert, allerdings fehlt die zeitliche Information wann diese Änderungen vorgenommen wurden. Um textgenetische Metainformationen auf Wort- bzw. Zeichenebene in das "ordered hierarchy of content objects model data" (OHCO) einer XML-Edition, wie das der BNE zu integrieren, schlägt das TEI-P5 Konsortium Fragmentierungs-, Milestone oder Standoff-Markup Annotationen vor (Jörg Hornschemeyer, 2013), die am WAB bisher nicht durchgeführt wurden. Von Geisteswissenschaftlern, deren wissenschaftliches Kerngebiet im Allgemeinen weit entfernt von der XML-Programmierung liegt, würde großer programmtechnischer Editionsaufwand verlangt. Eine Folge ist, dass von „Nachverwertern" der Edition zur Generierung der textlichen Varianten algorithmisches Ausmultiplizierten der annotierten Varianten implementiert wird, was z.B. in der Wittgenstein-Edition bei einzelnen Bemerkungen eine vierstellige Anzahl von Lesarten generiert. Betrachtet man die so automatisch generierten Lesarten, sind die meisten syntaktisch und semantisch falsch, was fatale Auswirkungen auf semantische Analysen der Textobjekte hat. Ohne zusätzliche, fein granulierte Metainformation

in den annotierten Varianten sind die Spielräume der automatisierten Lesartengenerierung jedoch nicht einzugrenzen.

Im Umfeld der Wittgensteinforschung gibt es eine Edition, die bis auf Zeichenebene zeitliche Informationen zur Textgenese liefert: Die Prototractatus-Tools (PTT 2016) von Martin Pilch (Pilch 2018). Sie dokumentieren den Nutzern Ludwig Wittgensteins Schreibprozess, beginnend mit einem leeren Notizbuch im Jahre 1915 und bis zum endgültigen Diktat des Ts-204 im Sommer 1918, das zu seiner einzigen philosophischen Veröffentlichung zu Lebzeiten, der „Logisch-Philosophischen Abhandlung“ führte. Leider konnten wir die Daten und Metainformationen der PTT-Edition in unserer FinderApp Infrastruktur nicht direkt analysieren, da unsere WiTTFind Infrastruktur zum einen auf das dokumentorientierte XML-TEI-P5 Datenformat aus Bergen zugeschnitten ist, und zum anderen die PTT-Edition im inkompatiblen Microsoft Word-97 Format vorliegt. Alle verfügbaren XML-TEI Importtools erfassten nur Bruchteile der Annotationen, sodass z.B. die Zeitinformationen der PTT überhaupt nicht erkannt und transformiert wurden. Um möglichst viel von der PTT-Textedition weiterzuverwenden, und damit der PTT-Hg. die Edition in seiner gewohnten Microsoft-Office Umgebung weiter optimieren kann, entwickelten wir eine mit Microsoft EXCEL leicht zu bedienende mehrdimensionale Tabellenstruktur. Die Editionsdaten und Metainformation der Word-97 Edition konnten wir größtenteils mit eigenen Programmen und Office-Macrotechniken transferieren. Zur Integration der Tabellen in die Infrastruktur unserer FinderApp verwendeten wir LibreOffice-Tools und selbst geschriebene Python Programme, die die Daten, sobald sie in das git-Repository des Projekts kopiert werden, mit Hilfe der continuous Integration automatisch transformieren und importieren. Zur Web-Präsentation werden sie an unsere neu entwickelte FinderApp, den OdysseeReader (siehe Abb. 1, odysseereader.wittfind.cis.lmu.de), übergeben. Dieses Vorgehen trennt zwar das Daten- und Repräsentationsmodell, jedoch entwickelten wir ein positionsinvariantes Siglensystem, bestehend aus dem Tupel (Zeitstempel, Dokument, Seite, Zeile, Zeichenposition), das die beiden Modelle eineindeutig verknüpft. Diese bijektive Relation zwischen den beiden Modellen definiert dem Hg., wo er in seinem Datenmodell Änderungen vornehmen muss um sie an eine bestimmte Stelle, zu einem bestimmten Zeitpunkt im Repräsentationsmodell zu platzieren.

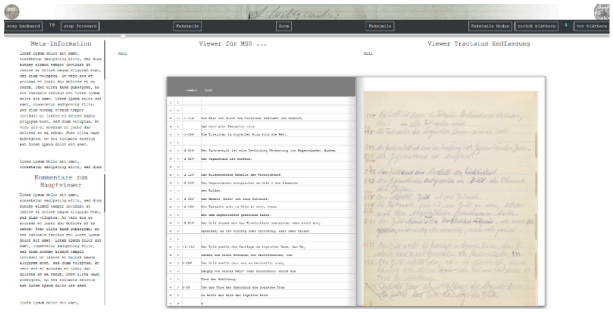

Abbildung 1: Der OdysseeReader odysseereader.wittfind.cis.lmu.de

Ähnlichkeitssuche mit vorgeschaltetem Semantic Clustering

Die Ähnlichkeitssuche WiTTSim berechnet mit Hilfe computerlinguistischer Methoden für jede Bemerkung einen „charakteristischen” Vektor, oder, intuitiv gesprochen: Man bestimmt einen “Fingerabdruck”. Dieser automatisierte Prozess wird unabhängig im Voraus berechnet, was spätere Prozesse vereinfacht und beschleunigt. Dieser „Fingerabdruck“ beinhaltet linguistische Informationen, wie beispielsweise Wörter, deutsche und englische Synonyme (aus Germanet und Wordnet), Wortarten (Treetagger) und Lemmata *(WiTTLex*, Röhrer 2019). Diese Informationen werden in binäre Vektoren übersetzt, welche insgesamt etwa 115.000 Features umfassen. Zusätzlich zur Datenbasis wurden 471 Bemerkungen bereits gruppiert, also mit Ground Truth Labels versehen. Die Gruppen bestehen dabei aus 2-15 Bemerkungen und das gelabelte das Korpus umfasst 1.670 Bemerkungen, was ca. 3% des gesamten Nachlasses entspricht.

Zur Semantischen Ähnlichkeitsberechnung ist allerdings eine Reduktion des Feature Raumes zwingend nötig, da die Vektoren mit so hoher Dimensionalität semantisch „weit voneinander entfernt“ sind und keine semantischen Gruppierungen auszumachen sind. Dieses Phänomen ist auch bekannt als *Curse of Dimensionality*. Daher werden die Vektoren zunächst auf eine angemessene Anzahl von Features skaliert, um sie anschließend clustern zu können. Verwendete Reduktionstechniken umfassen Singular Vector Decomposition (SVD), Principal Component Analysis (PCA), Sparse Random Projection (SRP) und Uniform Manifold Approximation and Projection (UMAP). Auf unseren Daten zeigte eine SVD Reduktion zu 1.600 Dimensionen die besten Ergebnisse, zusammen mit UMAP, welches darüber hinaus die Daten im zweidimensionalen Raum klar gruppiert. Letzteres erlaubt nur eine Zieldimension von 2 bis 100 Dimensionen, weshalb zum Erhalt der Varianz die maximale Dimensionsanzahl von 100 gewählt wurde, um

einen bestmöglichen Erhalt der gespeicherten Information zu gewährleisten.

Nach erfolgter Reduktion der Dimension können die Datenpunkte, also alle Bemerkungen, geclustert werden. Verwendete Clustering Techniken umfassen den klassischen K-Means Ansatz (Mac-Queen 1967, Ball and Hall 1956, Lloyd 1982, Steinhaus 1955), aber auch Dichte-basierte Ansätze wie Mean-Shift (Duda und Hart 1973) und DBSCAN (Ester et al. 1996), das statistische Gaussian Mixture Modell (Redner und Walker 1984) und das hierarchische Ward Clustering (Ward 1963). Beste Ergebnisse konnten mit einer Kombination von SVD und K-Means mit einer Anzahl an k=150 Clustern erzielt werden. Evaluiert wurde anhand der drei unüberwachten Metriken Silhouette Score, Davies Bouldin Index, und Calinski-Harabasz Index. Zusätzlich konnte durch die verfügbaren Ground Truth Labels auch der Recall berechnet werden, welcher in den Experimenten einen maximalen Wert von 1,0 erreicht. Dies zeigt, dass alle der gelabelten Daten richtig zugeordnet werden konnten. Wird eine Suchanfrage zum Auffinden ähnlicher Bemerkungen gestartet, muss nur der charakteristische Vektor der eingegebenen Bemerkung berechnet werden und das nächstliegende Cluster bestimmt werden. Letzteres erfolgt durch eine Bestimmung des am nächsten gelegenen Cluster Mittelpunkts (Zentroids). Anschließend werden die Abstände zu allen Bemerkungen des bestimmten Clusters gemessen, welche zuletzt dem Philologen zur genaueren Prüfung „gerankt“ vorgeschlagen werden.

Zusammenfassung und Ausblick

Unsere zeitgesteuerte textgenetische Edition kann von einem Wissenschaftler ohne XML Kenntnisse innerhalb einer Office Umgebung erstellt werden. Das continuous Integration System von git transferiert die Edition automatisch in unser WEB-basiertes Repräsentationssystem, den OdysseeReader. Über das von uns entwickelte eineindeutige Siglensystem verliert der Hg. niemals den klaren Zusammenhang zwischen Editions- und Präsentationsmodell.

Das von uns entwickelte Ähnlichkeitstool mit vorgeschaltetem Semantic Clustering könnte auch zur Ähnlichkeitsbestimmung zwischen zwei gegebenen Texten verwendet werden: Der Nutzer könnte einen Text eingeben, und es werden potentiell ähnliche Textpassagen in einer Sammlung von Texten gesucht, die dann „gerankt“ nach Ähnlichkeiten in einer Art Hitliste ausgegeben werden. Eine derartige Sortierung nach Textähnlichkeiten könnte es dem Philologen zum Beispiel besonders erleichtern, potentielle Zitate, Einflüsse und Verweise eines Autors innerhalb seines Werkes und im Bezug auf die Literatur seiner Zeit aufzuspüren.

Bibliographie

Ball, Geoffrey H. / Hall David J. (1965): *Isodata, a novel method of data analysis and pattern classification*. Technical report, Stanford research inst Menlo Park CA.

Duda, Richard O. / Hart, Peter E. (1973): "Pattern analysis and scene classification." *J. Wiley* 1:73.

Ester, Martin / Kriegel Hans-Peter / Sander, Jörg / Xu, Xiaowei et al. (1996): „A density-based algorithm for discovering clusters in large spatial databases with noise.“, in *KDD*, volume 96, pages 226–231.

Hadersbeck, Maximilian / Pichler, Alois / Fink, Florian / Gjesdal, Oyvind (2014): Wittgenstein’s Nachlass: WiTTFind and Wittgenstein Advanced Search Tools (WAST), DATeH, Madrid.

Hadersbeck, Maximilian / Still, Sebastian (2018): *Investigating Wittgenstein’s Nachlass: WiTTFind, WiTTReader, OdysseeReader and Wittgenstein Advanced Search Tools*, im Katalog zur Ausstellung „DIE TRACTATUS ODYSSEE“ S.127-137, Wittgenstein Initiative, Wien.

Lloyd, Stuart P. (1982): „Least squares quantization in pcm“, in: *IEEE transactions on information theory*, 28(2):129–137.

MacQueen, J. B. (1967): „Some methods for classification and analysis of multivariate observations.“, in: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967

Pichler, Alois / Krüger, Heinz W. / Smith, D. / Bruvik, Tone / Lindebjerg, Anne / Olstad, Vemund (Hrsg.) (2009): Wittgenstein Source Bergen Facsimile (BTE). Wittgenstein Source Bergen.

Redner, Richard A. / Walker, Homer F. (1984): Mixture densities, maximum likelihood and the em algorithm. SIAM review, 26(2):195–239.

Röhrer, Ines / Ullrich, Sabine / Hadersbeck, Maximilian (2019): *Weltkulturerbe international digital: Erweiterung der Wittgenstein Advanced Search Tools durch Semantisierung und neuronale maschinelle Übersetzung*. multimedial multimodal. Abstracts zur Jahrestagung des Verbandes Digital Humanities im deutschsprachigen Raum, 25. - 29.03.2019 an den Universitäten zu Mainz und Frankfurt.

Steinhaus, Hans (1955): Quelques applications des principes topologiques à la géométrie des corps convexes. Fund. Math, 41:284–290.

Ullrich, Sabine / Bruder, Daniel / Hadersbeck, Maximilian (2018): “Aufdecken von “versteckten” Einflüssen: Teil-Automatisierte Textgenetische Prozesse mit Methoden der Computerlinguistik und des Machine Learning”, 5. Tagung Digital Humanities im deutschsprachigen Raum 26.2.-2.3. (Köln).

Ullrich, Sabine (2019): *Boosting Performance of a Similarity Detection System using State of the Art Clustering Algorithms*. Master‘s thesis. LMU.

Pilch, Martin (2018): *Frontverläufe im Prototractatus – Zur gedanklichen Entwicklung von Krakau bis Sokal (1914/1915)*, Wittgenstein-Studien 9 (S.101-154), Internationale Ludwig Wittgenstein Gesellschaft (ILWG).

Still, Sebastian (2018): *Ludwig Wittgenstein: 100 Jahre Traktatus. Der Odyssee-Reader, ein web-basiertes Tool zur text-genetischen Suche im Traktatus*, Masterthesis, Ludwig-Maximilians-Universität München.

Feldweg, Birgit / Feldweg, Helmut (1997): „GermaNet - a Lexical-Semantic Net for German.", in: *Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid.

Henrich, Verena / Hinrichs, Erhard (2010): „GernEdiT - The GermaNet Editing Tool", in: *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*. Valletta, Malta, pp. 2228-2235.

Hörnschemeyer, Jörg / Thaller, Manfred / Förtsch, Reinhard (2017): *Textgenetische Prozesse in Digitalen Editionen*, Köln Universitäts- und Stadtbibliothek Köln 2017, https://www.worldcat.org/title/textgenetische-prozesse-in-digitalen-editionen/oclc/1002260195

Schmidt, Alfred (2018): „Ludwig Wittgenstein's Nachlass in the UNESCO Memory of the World register.", in: *Nordic Wittgenstein Review* 7(2):209–213.

UNESCO (2017): UNESCO-Weltdokumentenerbe - Zwei Neuaufnahmen. URL: https://www.unesco.at/presse/artikel/article/unesco-weltdokumentenerbe-zwei-neuaufnahmen/ [letzter Zugriff 19. Juni 2018].

Ward, John H. (1963): „Hierarchical grouping to optimize an objective function." *Journal of the American statistical association* 58.301: 236-244.