Weltkulturerbe international digital: Erweiterung der Wittgenstein Advanced Search Tools durch Semantisierung und neuronale maschinelle Übersetzung

Röhrer, Ines
i.roehrer@campus.lmu.de
LMU München, Deutschland

Ullrich, Sabine
sabine.ullrich@campus.lmu.de
LMU München, Deutschland

Hadersbeck, Maximilian
maximilian@cis.uni-muenchen.de
LMU München, Deutschland

Einleitung

Mit der Aufnahme des Nachlasses von Ludwig Wittgenstein ins Internationale UNESCO-Weltdokumentenregister im Jahr 2017, gewinnen die Forschung an den Werken des Philosophen, sowie die Texte selbst an großer Bedeutung (Trötzmüller 2017). Durch langjährige fachübergreifende Kooperation mit dem Wittgenstein Archiv der Universität Bergen (WAB) (Pichler 2010, 2014) kann das Centrum für Informations- und Sprachverarbeitung (CIS) der Ludwigs-Maximilians-Universität München einen umfassenden Zugang zum Nachlass Ludwig Wittgensteins anbieten. Der Zugang zum Nachlass wird mit einer Suchmaschine und integriertem Faksimile Reader über das Portal WiTTFind (http://wittfind.cis.uni-muenchen.de) ermöglicht. Zur Forschung an der textgenetischen Entwicklung des Prototraktatus wurde als neueste Applikation der Odyssee Reader entwickelt (Still 2018). Durch die intensive Zusammenarbeit mit den Philosophen konnte die Suchmaschine durch die Wittgenstein Advanced Search Tools (WAST) bereits umfangreich erweitert werden (Hadersbeck et al. 2012, 2014) und ermöglicht eine schnelle Suche von konkreten Textstellen im Nachlass, semantischen Ähnlichkeiten in den Themenbereichen Farbe (Krey 2014) und Musik (Röhrer 2017), einen Faksimile Reader (Lindinger 2015), sowie einen Geheimschrift-Übersetzer. Eine Übersicht inklusive der hier vorgestellten Erweiterungen ist in Abbildung 1 zu sehen.

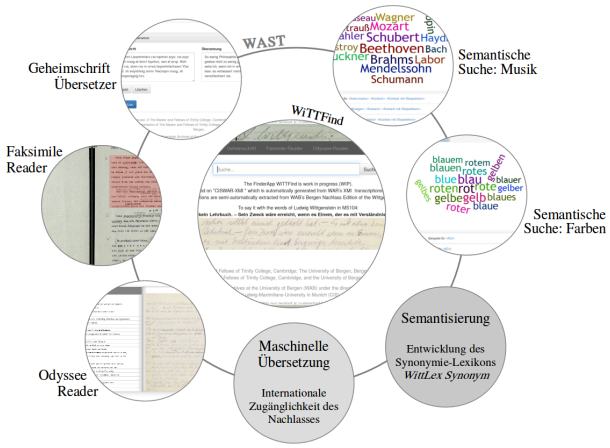

Abbildung 1. Übersicht der Wittgenstein Advanced Tools (WAST) in WiTTFind. Die grau hinterlegten Komponenten werden hier vorgestellt.

Integration der Semantisierung

Auf unserem Poster sollen nun zwei neue Komponenten der WAST vorgestellt werden. Zum einen wird das seit Jahren bestehende digitale Lexikon WiTTLex erweitert. Dieses Lexikon ist im DELA-Format verfasst und ermöglicht unserer FinderApp WiTTFind Suchbegriffe lemmatisiert und zerlegt nach Wortstamm und Partikel im Nachlass von Ludwig Wittgenstein zu finden. Die Besonderheit von WiTTLex besteht darin, dass es auf Wittgensteins Sprache zugeschnitten ist und nur Wörter enthält, die in seinem Nachlass vorkommen. Aufgrund dieser Eigenschaften bietet das Lexikon eine einzigartige Grundlage für sprachliche Untersuchungen in Ludwig Wittgensteins Werken. Um eine detailliertere Textforschung für Fragestellungen semantischer Natur zu ermöglichen wird derzeit im Rahmen eines studentischen Forschungsprojektes ein Synonymie-Speziallexikon, WiTTLex Synonym, entwickelt. Die Grundlage für dieses Lexikon ist einerseits die durch WiTTLex geschaffene Wortdatenbank, sowie andererseits die aus GermaNet (Hamp et al 1997, Henrich et al. 2010) und WordNet (Miller 1995, Fellbaum 1998) extrahierten Synonyme. Equivalent zu WorNet ist GermaNet ein lexikalisch-semantisches Wortnetzsystem, welches an der Universität Tübingen entwickelt wird. Anschließend wird diese Basis in eine dem DELA-System ähnliche Struktur formatiert, sowie manuell getestet und ergänzt. Das entstandene Lexikon kann die Suche von WiTTFind für die Nutzer anreichern, da ähnliche Textstellen gefunden werden können. Der Suchraum wird einerseits durch die Synonyme selbst, andererseits mit Wörtern erweitert, die eine Synonymverlinkung zum Suchwort haben. Ein derartiger schrittweiser Aufbau und

Erweiterung eines Synonymielexikons ermöglicht eine Evaluation von GermaNet im sprachlichen Kontext der Philosophie und kann zeigen, für wie viele Wörter Synonyme automatisch gefunden werden konnten, und von welcher Güte die gefundenen Synonyme sind. In einem zweiten Evaluationsverfahren wird verglichen, ob unser finales WiTTLex Synonym eine Verbesserung gegenüber einem rein automatischen, auf GermaNet und WordNet basierenden Systems, bei unserer Ähnlichkeitssuche WiTTSim auf Ludwig Wittgensteins Nachlass zeigt.

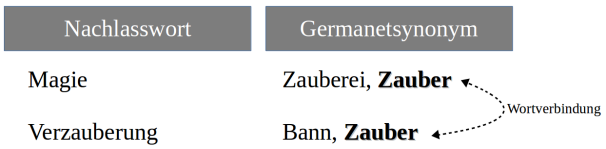

Abbildung 2. Beispiel für eine Synonymverlinkung

Neuronale maschinelle Übersetzung

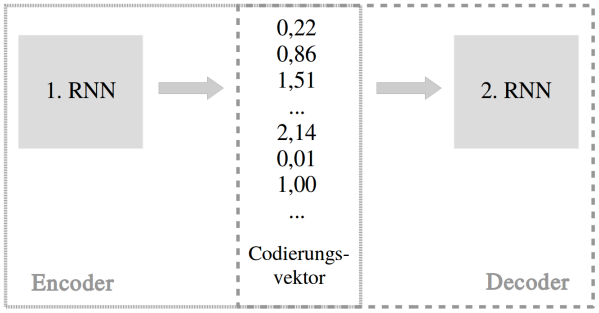

Abbildung 3. Übersicht des neuronalen maschinellen Übersetzungssystems.

Die zweite Erweiterung der WAST betrifft den internationalen Zugang der Suchmaschine für nicht deutschsprachige Wissenschaftler. Die gefundenen Faksimile und deren Transkription aus Bergen können derzeit nur in Originalsprache angezeigt werden. Nur ein sehr geringer Anteil dieser Originaltexte wurde auf Englisch verfasst, während der Großteil in deutscher Sprache geschrieben wurde. Daher wird derzeit im Rahmen eines weiteren studentischen Forschungsprojekts ein maschinelles Übersetzungssystem integriert, um Philosophen und anderen Interessierten aus aller Welt einen möglichst objektiven Zugang zum Nachlass zu ermöglichen. Das Übersetzungssystem wird als Sequence to Sequence Modell (Luong et al. 2015, 2017, Sutskever et al. 2014) implementiert. Dafür werden zwei rekurrente neuronale Netze (RNNs) trainiert, bestehend aus einem Encoder und einem Decoder. Der Encoder berechnet einen Codierungsvektor für den deutschen Textabschnitt, während der Decoder den entstandenen Vektor ins Englische transformiert (Abbildung 3). Werden diese zwei Netze aneinandergeschaltet, erhält man ein neuronales maschinelles Übersetzungssystem, welches den Nachlass vom Deutschen ins Englische übersetzt.

Es muss jedoch angemerkt werden, dass die automatische Übersetzung keinesfalls eine philosophisch-interpretatorische Übersetzung ersetzen kann. Sie kann lediglich eine Grundlage bilden, welche dann in intensiver Zusammenarbeit mit den Philosophen geprüft und optimiert werden kann.

Für zukünftige Arbeiten sollen die übersetzen Texte in die Ähnlichkeitssuche WiTTSim einfließen (Ullrich et al. 2018), wo sie der Aufdeckung von sprachübergreifende Ähnlichkeiten dienen kann (siehe Abbildung 4).

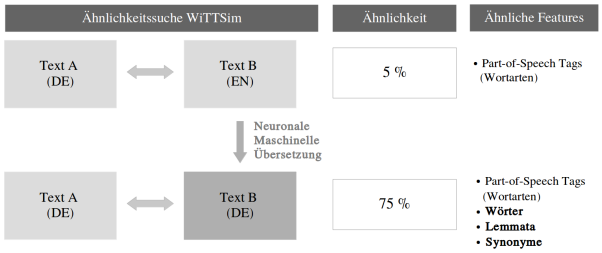

Abbildung 4. Ähnlichkeitsberechnung mit WiTTSim und Vergleich der Ergebnisse mit und ohne maschineller Übersetzung am Beispiel von zwei Texten A und B.

Bibliographie

Fellbaum, Christiane (1998): *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Hadersbeck, Maximilian / Pichler, Alois / Fink, Florian / Seebauer, Patrick / Strutynska, Olga (2012): *New (re)search possibilities for Wittgenstein's Nachlass*. 35th International Wittgenstein Symposium 2012, Kirchberg am Wechsel, Contributions, pp. 102-105. Kirchberg am Wechsel: ALWS.

Hadersbeck, Maximilian / Pichler, Alois / Fink, Florian / Gjesdal, Øyvind (2014): *Wittgenstein's Nachlass: WiTTFind and Wittgenstein advanced search tools (WAST)*. Digital Access to Textual Cultural Heritage 2014 (DaTeCH 2014), pp. 91-96. Madrid.

Hadersbeck, Maximilian / Pichler, Alois / Fink, Florian / Bruder, Daniel / Arends, Ina / Baiter, Johannes (2015): *Wittgensteins Nachlass: Erkenntnisse und Weiterentwicklung der FinderApp WiTTFind*. 2. Tagung Digital Humanities im deutschsprachigen Raum 23.-27.2 (Graz).

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

Henrich, Verena / Erhard Hinrichs (2010): *GernEdiT - The GermaNet Editing Tool*. Proceedings of the Seventh

Conference on International Language Resources and Evaluation (LREC 2010). Valletta, Malta, pp. 2228-2235.

Krey, Angela (2014): *Semantische Annotation von Adjektiven im Big Typescript von Ludwig Wittgenstein*. Bachelorarbeit, CIS.

Lindinger, Matthias (2015): *Entwicklung eines WEB-basierten Faksimileviewers mit Highlighting von Suchmaschinen-Treffern und Anzeige der zugehörigen Texte in unterschiedlichen Editionsformaten*. Masterthesis, CIS.

Luong, Minh-Thang / Hieu Pham / Christopher D Manning (2015): *Effective approaches to attention-based neural machine translation* EMNLP.

Luong, Minh-Thang / Eugene Brevdo / Rui Zhao (2017): *Neural Machine Translation (seq2seq) Tutorial*, https//github.com/tensorflow/nmt, zugegriffen am 9.10.2018.

Miller, George A. (1995): *WordNet: A Lexical Database for English*, in: Communications of the ACM Vol. 38, No. 11: 39-41.

Pichler, Alois (2010): *Towards the New Bergen Electronic Edition*, in: Wittgenstein After His Nachlass. Ed. Nuno Venturinha, pp. 157-172. Houndmills: Palgrave Macmillan.

Pichler, Alois / Bruvik, Tone Merete (2014): *Digital Critical Editing: Separating Encoding from Presentation*, in: Digital Critical Editions. Ed. Daniel Apollon, Claire Bélisle, Philippe Régnier, pp. 179-202. Urbana Champaign: University of Illinois Press.

Röhrer, Ines (2017): *Musik und Ludwig Wittgenstein: Semantische Suche in seinem Nachlass*, Bachelorarbeit, CIS.

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.

Sutskever, Ilya / Oriol Vinyals, / Quoc V. Le. (2014): *Sequence to sequence learning with neural networks*, NIPS.

Trötzmüller, Eva (2017): *Unesco-Weltdokumentenerbe - Zwei Neuaufnahmen*, https://www.unesco.at/presse/artikel/article/unesco-weltdokumentenerbe-zwei-neuaufnahmen/, zugegriffen am 12.10.2018

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)