Tool zur Normalisierung und Historisierung

Eder, Elisabeth
e_eder@gmx.net
Ludwig Maximilians Universität München, Deutschland

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

Das in diesem Poster vorgestellte, unter ‹http://goethefind.cis.uni-muenchen.de/?translator› verfügbare Translationstool überführt historisches Deutsch aus einem ungefähren Zeitraum von 1750 bis 1850 in gegenwartssprachliches Deutsch und umgekehrt modernen deutschen Text in seine historische Version.

Für eine Normalisierung oder Modernisierung von historischen Wörtern wurden in den letzten Jahren unterschiedliche Herangehensweisen präsentiert. Neben einer Modernisierung über Lexikon-Lookup, Transkriptionsregeln, Levenshtein-Distanz oder phonologische Ähnlichkeit fanden auch Methoden der statistischen maschinellen Übersetzung Anwendung (Scherrer / Erjavec 2015: 2f.). Um orthographischen Differenzen bei einer Translation einzelner Wörter aus eng verwandten Sprachen gerecht zu werden, werden dabei im Gegensatz zur standardmäßigen phrasenbasierten statistischen maschinellen Übersetzung die Phrasen nicht aus Wörtern, sondern aus Buchstabensequenzen gebildet und anstelle der Wörter der Ausgangs- und der Zielsprache die Buchstaben der Wortpaare aligniert [Pettersson et al., 2014]. Buchstabenbasierte statistische maschinelle Übersetzung zur Normalisierung historischer Wörter wurde vielfach mit dem Tool ‹Moses› (Koehn et al. 2007) durchgeführt, wie beispielsweise bei (Pettersson et al. 2014), (Nakov / Tiedemann 2012) oder (Scherrer / Erjavec 2015). Neben einem Gebrauch zur Normalisierung wird dieses hier auch für die umgekehrte Überführungsrichtung eingesetzt.

In einem weiteren Ansatz zur Modernisierung und Historisierung bedient sich das Translationstool zudem neuronaler maschineller Übersetzung. Das dabei häufig verwendete Encoder-Decoder-Modell übertragen Faruqui, Tsvetkov, Neubig und Dyer (2016) auf die buchstabenbasierte Generierung von Wortflexion. Aufgrund der ähnlichen Grundlage kommt deren Tool ‹Morph-Trans› , das sich aus LSTMs, einer speziellen Form von rekurrenten neuronalen Netzen, zusammensetzt, zum Einsatz. Nach Wissen der Autoren ist dies der erste Versuch, ein neuronales Encoder-Decoder-Modell für eine Historisierung und Normalisierung deutscher Texte zu gebrauchen.

Als Trainings- und Entwicklungsdatensätze für die beiden Methoden dienten Wörter von 200 literarischen Texten aus einem Zeitraum von 1749 bis 1850. Diese Wörter wurden mithilfe des ‹Cascaded Analysis Broker› vom Deutschen Textarchiv normalisiert, um im Anschluss daran auf die derzeit gültige ‹s›-Schreibung aktualisiert zu werden. Aus den historischen und den modernen Schreibweisen der Wörter wurden das Grundkorpus sowie ein Lookup-Lexikon gebildet. Im Translationstool werden die beiden Ansätze zusätzlich auch in Kombination mit diesem Lexikon eingesetzt. Zu Vergleichszwecken sind diese vier unterschiedlichen Ausgaben des Weiteren um ein auf einfachen Überführungsregeln und regulären Ausdrücken basierendes Verfahren ergänzt. Die unterschiedlichen Herangehensweisen können online anhand eigener Beispiele gegenübergestellt werden.

Tests auf exemplarischen Datensätzen zeigten, dass buchstabenbasierte statistische maschinelle Übersetzung nicht nur für eine Modernisierung, sondern im Deutschen ebenso für eine Historisierung dienlich ist und auch das neuronale Encoder-Decoder-Modell im Hinblick auf beide Überführungsrichtungen nutzbringend eingesetzt werden kann, wobei das Normalisieren im Vergleich zum Historisieren, wie zu erwarten war, durchweg bessere Resultate erzielte, was wohl unter anderem der Fülle an orthographischen Varianten in historischen Texten geschuldet ist.

Im geisteswissenschaftlichen Kontext ist eine Modernisierung historischer Wörter oftmals für eine erfolgreiche Anwendung sprachtechnologischer Werkzeuge, wie zum Beispiel Part-of-Speech-Tagger, auf älteren Texten von Nöten, während eine Historisierung beispielsweise bei der Suche auf historischem Text zu einer erheblichen Erleichterung beitragen könnte, indem der moderne Suchterm historisiert wird, da von Anwendern und Anwenderinnen nicht erwartet werden kann, dass sie um die alten Schreibweisen der Wörter wissen. Eine Verwendung von buchstabenbasierter statistischer maschineller Übersetzung und buchstabenbasierten neuronalen Encoder-Decoder-Modellen zur Normalisierung und Historisierung bezüglich solcher Aufgaben und ähnlichen Problemstellungen im Bereich der Geisteswissenschaften ist vorstellbar.

Fußnoten

1. Online verfügbar unter: http://www.statmt.org/moses/
2. Online verfügbar unter: https://github.com/mfaruqui/morph-trans
3. Online verfügbar unter: http://www.deutschestextarchiv.de/demo/cab/

Bibliographie

Faruqui, Manaal / Tsvetkov, Yulia / Neubig, Graham / Dyer, Chris (2016): „Morphological Inflection

Generation Using Character Sequence to Sequence Learning“, in: *Proceedings of NAACL* http://arxiv.org/pdf/1512.06110.pdf [letzter Zugriff 1. August 2016].

Koehn, Philipp / Hoang, Hieu / Birch, Alexandra / Callison-Burch, Chris / Federico, Marcello / Bertoldi, Nicola / Cowan, Brooke / Shen, Wade / Moran, Christine / Zens, Richard / Dyer, Chris / Bojar, Ondrej / Constantin, Alexandra / Herbst, Evan (2007): „Moses: Open Source Toolkit for Statistical Machine Translation“, in: *Annual Meeting of the Association for Computational Linguistics (ACL)*, demonstration session, Prague, Czech Republic.

Nakov, Preslav / Tiedemann, Jörg (2012): „Combining Word Level and Character-Level Models for Machine Translation Between Closely Related Languages“, in: *Proceedings of ACL-2012*

Pettersson, Eva / Megyesi, Beáta / Nivre, Joakim (2014): „A Multilingual Evaluation of Three Spelling Normalisation Methods for Historical Text“, in: *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH) @ EACL 2014* 32–41.

Scherrer, Yves / Erjavec, Tomaž (2015): „Modernising historical Slovene words“, in: *Natural Language Engineering* http://archiveouverte.unige.ch/unige:82305 [letzter Zugriff 1. August 2016]