SE Digitale Methoden

Semesterprojektbeschreibung
Normalisierung historischer Texte (mit neuronalen Netzen)

Wenyuan

31. März 2021

1 Einführung

Die Projektidee stammt aus der kürzlich veröffentlichten Arbeit "Semi-supervised Contextual Historical Text Normalization" [Makarov and Clematide, 2020] vom Institut für Computerlinguistik der UZH. Sie schlugen einen neuartigen Ansatz zur automatischen Normalisierung historischer Texte in die Standardvarietät vor und erreichten in realistischen Trainingsszenarien die gleiche Genauigkeit wie die manuelle Textnormalisierung. Das Korpus, das sie für die Experimente verwendeten, besteht aus mehreren Sprachen, und der deutsche Teil davon stammt hauptsächlich aus dem 14. bis 16. Jahrhundert. Eine mögliche Richtung des Projekts könnte die Evaluierung und Ausweitung des Ansatzes auf ältere Sprachen sein, z.B. Kymrisch, Bretonisch.

2 Daten (Korpus)

Die folgende Tabelle zeigt die Details der Datensätze, die in dem von Makarov und Clematide vorgeschlagenen Ansatz verwendet werden.

Language	Source Corpus	Time Period	Genre	Tokens (total)	Source of Splits
English	[ICAMET]	1386-1698	Letters	188,158	[HistCorp]
German	[Anselm]	14th-16th c.	Religion	71,570	prev. unpublished
German	[RIDGES]	1482-1652	Science	71,570	prev. unpublished
Hungarian	[HGDS]	1440-1541	Religion	172,064	[HistCorp]
Icelandic	[IcePaHC]	15th c.	Religion	65,267	[HistCorp]
Portuguese	[Post Scriptum]	15th-19th c.	Letters	306,946	prev. unpublished
Slovene	[goo300k]	1750-1899	Mixed	326,538	[KonvNormSl 1.0]
Spanish	[Post Scriptum]	15th-19th c.	Letters	132,248	prev. unpublished
Swedish	[GaW]	1527-1812	Official Records	65,571	[HistCorp]

3 Aufgabe

Eine mögliche Projektausrichtung könnte die Erweiterung des Ansatzes auf die Sprachen wie Kymrisch und Bretonisch, wie von Paul Widmer vorgeschlagen. Allerdings wäre es sehr aufwändig, das dafür benötigte Parallelkorpus zu erstellen, d.h. historische Texte und entsprechende Standardtexte als Gold Standard manuell zu normalisieren. Daher habe ich den Startpunkt des Projekts angepasst:

- einer Liste von Paaren moderner und historischer Wörter (z.B. ein paar Hundert bis Tausend) sammeln und verschiedener Normalisierungsmodelle zur historischen Textnormalisierung ausprobieren GitHub
- einen Vergleich mit dem von [Makarov and Clematide, 2020] vorgeschlagenen hauseigenen neuronalen Ansatz machen, der mit wenigen Trainingsdaten gut funktioniert GitHub
- (optional) die historische Textnormalisierung über den neuronalen Ansatz durchführen und einen menschlichen SprachwissenschaftlerIn, diese manuell zu korrigieren bitten, um sie auszuwerten.

4 Ressourcen

Ressourcen, die für das Projekt benötigt werden:

- historische Texte (in digitaler Form)
- Paare von modernen und historischen Wörtern
- einen Linguisten, der die verallgemeinerten modernen Texte begründen und bewerten kann

Sprachen:

- Walisisch
- Bretonisch

5 Related Work

Die oben genannten Ressourcen wurden ursprünglich zusammen mit [Bollmann, 2019] veröffentlicht, außerdem bietet [Bollmann, 2018] weitere Details und Hintergrundinformationen. [Korchagina, 2017] testete und evaluierte die folgenden drei Ansätze zur Textkanonisierung an historischen deutschen Texten aus dem 15. bis 16. Jahrhundert: regelbasiert, statistische maschinelle Übersetzung und neuronale maschinelle Übersetzung (NMT). Während [Hämäläinen et al., 2019] die NMT-Methoden überprüfte und verschiedene Methoden zur Verbesserung des Normalisierungsprozesses diskutierte. [Makarov and Clematide, 2020] entwickelte neue Ansätze zur semi-supervised kontextualisierten Textnormalisierung.

Literatur

[Bollmann, 2018] Bollmann, M. (2018). *Normalization of historical texts with neural network models.* doctoralthesis, Ruhr-Universität Bochum, Universitätsbibliothek.

[Bollmann, 2019] Bollmann, M. (2019). A large-scale comparison of historical text normalization systems. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3885–3898, Minneapolis, Minnesota. Association for Computational Linguistics.

[Hämäläinen et al., 2019] Hämäläinen, M., Säily, T., Rueter, J., Tiedemann, J., and Mäkelä, E. (2019). Revisiting NMT for normalization of early English letters. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 71–75, Minneapolis, USA. Association for Computational Linguistics.

[Korchagina, 2017] Korchagina, N. (2017). Normalizing medieval German texts: from rules to deep learning. In *Proceedings of the NoDaLiDa 2017 Workshop on Processing Historical Language*, pages 12–17, Gothenburg. Linköping University Electronic Press.

[Makarov and Clematide, 2020] Makarov, P. and Clematide, S. (2020). Semi-supervised contextual historical text normalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7284–7295, Online. Association for Computational Linguistics.