Workshop: Generative KI, LLMs und GPT bei digitalen Editionen
DHd, Passau, 27.02.2024
Yannic Bracke, yannic.bracke@bbaw.de
Berlin-Brandenburgische Akademie der Wissenschaften, Projekt: Text+

LLM-basierte Normalisierung historischer Schreibweisen mit *transnormer*

Normalisierung historischer Schreibweisen

Übertragung eines historischen Texts in moderne Rechtschreibung

Sie giengen beyde in dem kǒniglichen Spatzierhofe auff vnd nider.

Sie gingen beide in dem königlichen Spazierhof auf und nieder.

(Barclay 1626)

Wozu normalisieren?

- Besser lesbar
- Besser **durchsuchbar**
- Erschließt den Text für die Nutzung von **NLP-Tools**

Sie giengen beyde in dem kǒniglichen Spatzierhofe auff vnd nider.

Sie gingen beide in dem königlichen Spazierhof auf und nieder.

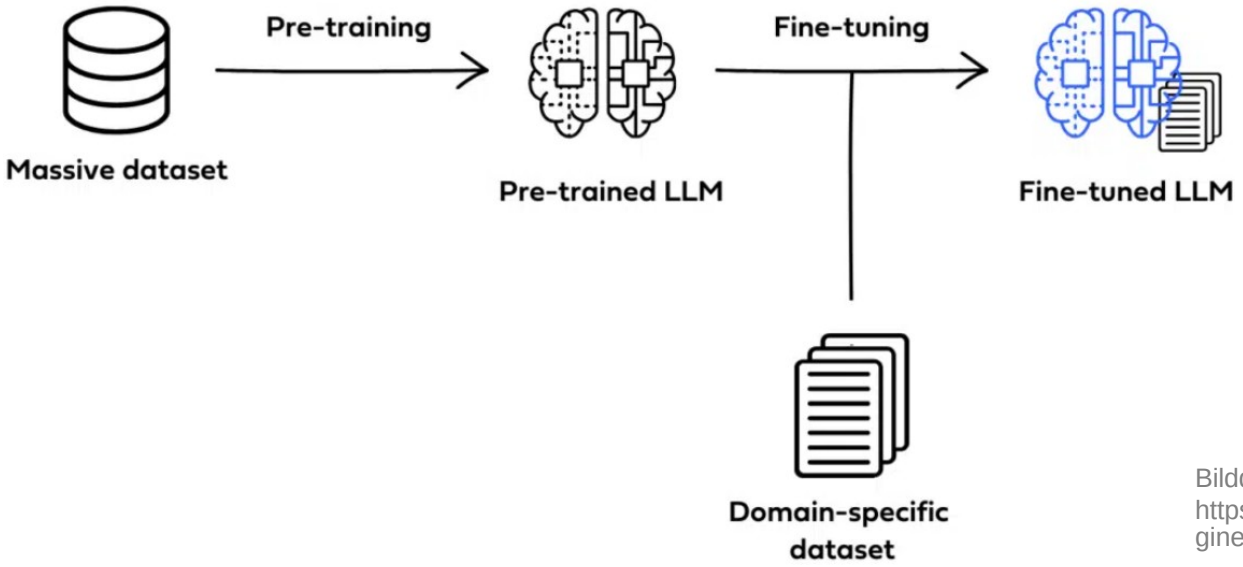
(Barclay 1626)

Automatische Normalisierung

- Mehrere Optionen: Von Ersetzungslisten bis Machine Learning
- Etabliertes Tool: *Cascaded Analysis Broker*, kurz*:* ***CAB*** (Jurish 2011)
  - Normalisierung des **DTA** und Einsatz in vielen weiteren Projekten
  - Webservice: https://www.deutschestextarchiv.de/public/cab/
- In der Entwicklung: ***transnormer***

transnormer: Grundidee

Normalisierung ~ maschinelle Übersetzung (vgl. Bollmann 2018)



Bildquelle: Najeeb Nabwani, https://deci.ai/blog/fine-tuning-peft-prompt-engineering-and-rag-which-one-is-right-for-you/

transnormer: Trainingsdaten fürs Finetuning

Korpus	Zeitraum	Tokens	Normalisierung
DTA EvalCorpus	*1780-1901*	*5M*	*manuell geprüft*
DTA Kernkorpus	*1598-1913*	*150M*	*CAB + Korrektur*
GerManC-GS	*1659-1786*	*60.000*	*manuell*
RIDGES	*1482-1887*	*300.000*	*manuell*

Können Sie weitere **manuell normalisierte Texte** zur Verfügung stellen? Sprechen Sie mich gerne an.

transnormer: Implementierung

- Basismodell: Encoder-Decoder *ByT5* (Xue et al. 2022; HuggingFace)
- Training/Evaluation mit Python und transformers-Bibliothek
- GitHub-Projektseite: *https://github.com/ybracke/transnormer*

transnormer: Demo

Demo (work-in-progress!):
https://transnormer:fah9Thai@riker.bbaw.de/transnormer/

Vielen Dank an Gregor Middell für die Bereitstellung des Webservice!

transnormer: Evaluation

- Evaluation für **DTA EvalCorpus** (1780-1901)
- Metrik: Word accuracy (Bawden et al. 2022)

Methode	**Accuracy**	**Accuracy (uncased)**
Identität	*79.59*	*79.80*
Transliteration	*93.91*	*94.17*
transnormer	***98.93***	***99.18***

transnormer: Wie geht es weiter?

Geplant:

- **Verbesserung** und **Erweiterung** der Trainingsdaten

  ➔ Training neuer Modelle
- Modelle für verschiedene historische **Zeiträume**
- **Veröffentlichung** von Modellen auf Huggingface
- Ausbau des **Webservices**
- Integration in ***MONAPipe*** (Dönicke et al. 2022)

Vielen Dank für die Aufmerksamkeit!

Quelle historischer Beleg

Barclay, John (1626). Johann Barclaÿens Argenis Deutsch gemacht durch Martin Opitzen. Breslau. In: Deutsches Textarchiv <https://www.deutschestextarchiv.de/barclay_argenis_1626>

Literatur

Bawden, R., Poinhos, J., Kogkitsidou, E., Gambette, P., Sagot, B., & Gabay, S. (2022). Automatic Normalisation of Early Modern French. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, 3354–3366. https://aclanthology.org/2022.lrec-1.358

Bollmann, M. (2018). Normalization of historical texts with neural network models. Dissertation. Ruhr-Universität Bochum. https://doi.org/10.13154/294-6213

Dönicke, T., Barth, F., Varachkina, H., & Sporleder, C. (2022). MONAPipe: Modes of Narration and Attribution Pipeline for German Computational Literary Studies and Language Analysis in spaCy. In Proceedings of the 18th Conference on Natural Language Processing (KONVENS 2022), 8–15. https://aclanthology.org/2022.konvens-1.2

Jurish, B. (2011). Finite-State Canonicalization Techniques for Historical German. Dissertation. Universität Potsdam. https://publishup.uni-potsdam.de/frontdoor/index/index/docId/5562

Xue, L., Barua, A., Constant, N., Al-Rfou, R., Narang, S., Kale, M., Roberts, A., & Raffel, C. (2021). ByT5: Towards a token-free future with pre-trained byte-to-byte models. https://doi.org/10.48550/ARXIV.2105.13626