

Historische Textnormalisierung

Herausforderungen und Potentiale von Deep Learning

Yannic Bracke (Text+, BBAW)
Anton Ehrmanntraut (JMU Würzburg)

DHd2025 – 06.03.2025

Historische Schreibungen

Das **Hei**ßeste liegt der Leidenschaft zu
aller**nä**chst, bemerkte Charlotte. Lehne, **so**
lange es noch Zeit **ist**, den guten **Rath** nicht
ab, nicht die **Hülfe** die ich uns biete. In
trüben Fällen muß derjenige wirken und hel
fen der am **klärsten** sieht. Dießmal bin

- Begegnen uns in vielen DH-Kontexten
- Historische Schreibungen weichen von heutiger Schreibung ab
- Gründe: Entstehung bzw. Wandel von Konventionen, Variation

Probleme

Orthographische Abweichungen erschweren:

- Volltextsuche (Bsp.: *heirathen*, *heyrathen*, *heurathen*, *heiraten*)
- Weiterverarbeitung mit NLP-Tools
- *Distant Reading* Ansätze

Textnormalisierung

Das **Äußerste** liegt der Leidenschaft zu allernächst, bemerkte Charlotte. Lehne, **so lange** es noch Zeit ist, den guten **Rath** nicht ab, nicht die **Hülfe** die ich uns biete. In trüben Fällen muß derjenige wirken und helfen der am **klarsten** sieht. Dießmal bin

Das **Äußerste** liegt der Leidenschaft zu allernächst, bemerkte Charlotte. Lehne, **solange** es noch Zeit ist, den guten **Rat** nicht ab, nicht die **Hilfe** die ich uns biete. In trüben Fällen muss derjenige wirken und helfen der am **klarsten** sieht.

Wie soll der normalisierte Text aussehen?

Ich wil mir in meinem Sterben fürstellen dich o Jesu!

Wie normalisieren bei

- veränderter Wortbildung?
- abweichender Wortstellung?
- morphosyntaktischen Abweichungen?
- ausgestorbenen Lemmata?

Welche Relation zwischen
historischen und normalisierten
Tokens? (1:1, 1:n, n:1, n:m)

In historischen Korpusprojekten ist Normalisierung oft in Guidelines festgelegt.
(Durrell et al., 2012; Krasselt et al., 2015; Odebrecht et al., 2020)

Automatische Normalisierung: Verschiedene Ansätze

Regel-basiert

- CAB (Jurish, 2012)
- Formale Regeln für Zeichen-ersetzungen im historischen Text
- Methode: endliche Automaten

DTA::CAB Web Service v1.115

Query: Lehne den guten Rath nicht ab, nicht die Hülfe die ich un:

Analyzer: default

Format: CSV (TAB-separated)

Flags: ☒ pretty ☒ clean ☒ exlex

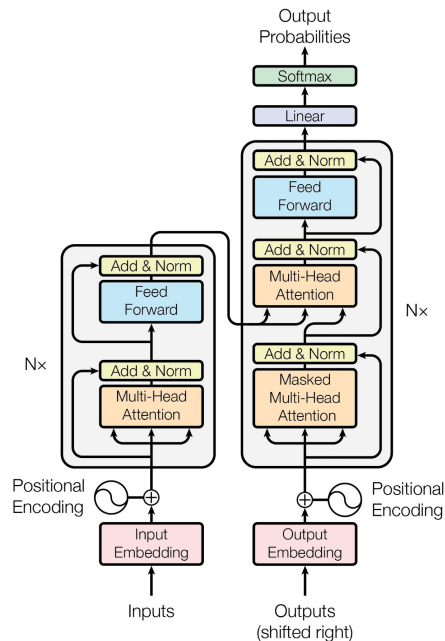
Options: {}

URL: <https://deustextarchiv.de/public/cab/query?a=default&fmt=csv&clean=1&pretty=1&raw=1&q=Lehne%20den%20guten%20f>

```
%% $:lang=de
Lehne  Lehne  Lehne  NN      Lehne
den    den    den    ART     d
guten  guten  guten  ADJA    gut
Rath   Rath   Rat    NN      Rat
nicht  nicht  nicht  PTKNEG nicht
ab     ab     ab     PTKVZ   ab
,      ,      ,      $,      ,
nicht  nicht  nicht  PTKNEG nicht
die    die    die    ART     d
Hülfe  Hülfe  Hilfe  NN      Hilfe
die    die    die    ART     d
ich    ich    ich    PPER    ich
uns    uns    uns    PPER    wir
biete  biete  biete  VVFIN   bieten
.      .      .      $.      .
```

Machine Learning

- Unsere Ansätze
- Modelle lernen statistische Zusammenhänge aus annotierten Daten (hier: Parallelkorpus)
- Methode: Transformer



Das *DTA-Eval*-Parallelkorpus (Jurish et al., 2013)

- Wort-aligniertes Parallelkorpus zwischen digitalen DTA-Editionen von Erstdrucken und gegenwärtigen Editionen aus TextGrid
- Algorithmisch bestimmt
+ händisch korrigiert
- Sample von 85 (belletristischen) Dokumenten, ca. 19. Jhdt.
→ 188 000 Sätze, 4 Millionen Tokens
- Keine konkreten Guidelines, nur die (implizite) Normalisierungspraxis der TextGrid-Editionen

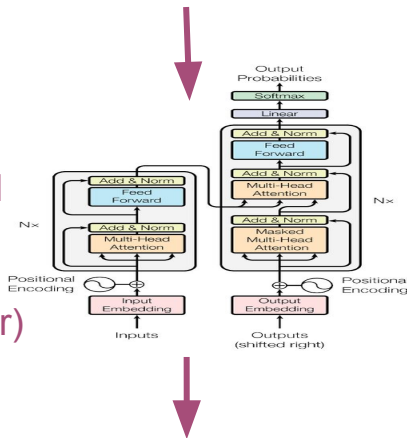
```
<s>
  <w class="LEX" new="Mir" old="Mir"/>
  <w class="SPLIT" new="widert es" old="widert's"/>
  <w class="LEX" new="," old=","/>
  <w class="LEX" new="die" old="die"/>
  <w class="LEX" new="Verworrenen" old="Verworrenen"/>
  <w class="JOIN" new="Dahinzuschlachten" old="Dahin zu fchlachten">
    <w class="JOIN" new="Dahin" old="Dahin"/>
    <w class="JOIN" new="zu" old="zu"/>
    <w class="JOIN" new="schlachten" old="fchlachten"/>
  </w>
  <w class="LEX" new="," old=","/>
  <w class="LEX" new="ihrer" old="ihrer"/>
  <w class="LEX" new="Torheit" old="Thorheit"/>
  <w class="LEX" new="Opfer" old="Opfer"/>
  <w class="LEX" new="." old="."/>
</s>
```

Wie mittels Transformern Texte normalisieren?

Satz-Ebene: Normalisiere Satz für Satz
("Transnormer") [à la text translation]

"Der Officier mußte sich dazu setzen und
ließ sich's wohl feyn"

fine-tuned
ByT5
(300M
Parameter)



"Der Offizier musste sich dazu setzen und
ließ sich's wohl sein"

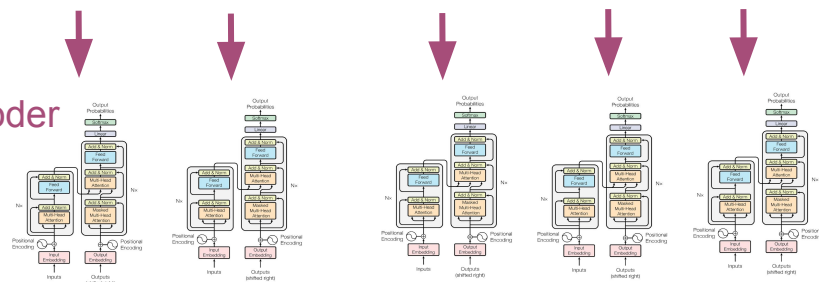
1. encoder-decoder
(8M Parameter)

2. Normalisierungs-
hypothesen

3. Ranking
mit GPT2
(130M Parameter)

Wort-Ebene: Normalisiere Wort für Wort

"Der" "Officier" "mußte" ... "wohl" "feyn"



"Der" "Offizier" "musste" ... "wol" "sein"
"Offizier" "wohl" "seien"

■ "Der Offizier musste ... wol sein"

■ "Der Offizier musste ... wohl sein"

■ "Der Offizier musste ... wohl seien"

Wie mittels Transformern Texte normalisieren?

Satz-Ebene: Normalisiere Satz für Satz
("Transormer")

- ✓ einfaches Setup / Training
(wie translation model)
- ✓ Benötigt keinen Tokenizer,
Flexibilität bei Worttrennung,
Satzbau
- ✗ Mehr Rechenleistung nötig

Wort-Ebene: Normalisiere Wort für Wort

- ✗ Konzeptionell aufwändiger
im Aufbau
- ✗ Satzbau fest, Wortrennung
nur über "Kontrollzeichen"
umsetzbar
- ✓ Sehr kleines Modell, läuft
auch auf CPU schnell

Ergebnisse (separates Test-Korpus, 16 Dokumente, 600k Tokens)



3.49% Fehler bei Beibehalten d. Originals

1.93% Fehler bei CAB (Jurish, 2012)

1.07% Fehler bei Csmtiser (Ljubešić et al., 2016)

0.84% Fehler bei ByT5 auf Satz-Ebene

0.81% Fehler bei Encoder-Decoder auf Wort-Ebene +LM

Ergebnisse (separates Test-Korpus, 16 Dokumente, 600k Tokens)

3.49% Fehler bei Beibehalten d. Originals

16.09% OOV

1.93% Fehler bei CAB (Jurish, 2012)

[Kein Training: OOV nicht berechenbar]

1.07% Fehler bei Csmtiser (Ljubešić et al., 2016)

10.84% OOV

0.84% Fehler bei ByT5 auf Satz-Ebene

7.76% OOV

0.81% Fehler bei Encoder-Decoder auf Wort-Ebene +LM

8.30% OOV

OOV =
hist. Token nicht im
Trainings-Datensatz

Nutzung

✓ Einbindung des Normalisierungsmodells in eigene lokale Workflows

✓ (Nach)Trainieren möglich

✗ Höhere Hardwareanforderungen und Ressourcenverbrauch
→ aber Inferenz ist auch ohne GPU möglich

```
from transformers import pipeline

transnormer = pipeline(model='ybracke/transnormer-19c-beta-v02')
sentence = "Die Königin faß auf des Pallaftes mittlerer Tribune."
print(transnormer(sentence))
# >>> [{'generated_text': 'Die Königin saß auf des Palastes mittlerer Tribüne.'}]
```

[Satz-Ebene]

Transnormer (fine-tuned ByT5)



Hugging Face

Search models, datasets, users...

ybracke/transnormer-19c-beta-v02

like 0

ybracke/transnormer-18-19c-beta-v01

private

[Wort-Ebene]

Encoder-Decoder + LM

aehrm/
hybrid_textnorm

Text normalization with hybrid model architecture



1

Contributor



0

Issues



1

Star



0

Forks



```
bash $ pip  
bash $ nor
```

Datasets: ybracke/dta-reviEvalCorpus-v1

Tasks: Text2Text Generation Modalities: Text Formats: json

Size: 100K

Datasets: ybracke/dtak-transnormer-basic-v1

Tasks: Text2Text Generation Modalities: Tabular Text Format

Languages: German Size: 1M - 10M Libraries: Datasets Dask

License: cc-by-sa-4.0

Dataset card Viewer Files Community

Dataset Viewer

Auto-converted to Parquet API Embed

Historical German Text Normalization Using Type- and Token-Based Language Modeling

Anton Ehrmanntraut

Julius-Maximilians-Universität Würzburg
anton.ehrmanntraut@uni-wuerzburg.de

September 5, 2024

arXiv: 2409.02841

Abstract

Historic variations of spelling poses a challenge for full-text search or natural language

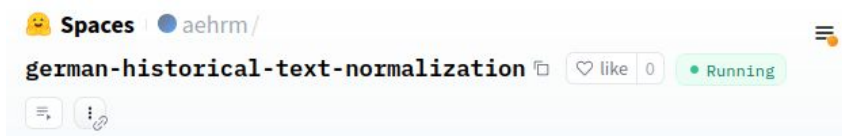
One major obstacle to this research design is language change. The older the historical texts are, the more they diachronically deviate from the current standard language in orthography, syntax, lex-

Ausblick

- Veröffentlichung von Transnormer-Modell für Zeitraum ab 1600
- Planungen: Deployment von Transnormer-Modellen als Text+-Service
- Kontakt: **textplus@bbaw.de**

For the meantime:

huggingface.co/spaces/aeherm/german-historical-text-normalization



German Historical Text Normalization

Input

Model

ybracke/transnormer-19c-beta-v02 (fast) ▾

Clear

Submit

Output

Vielen Dank für die
~~auffmerksamkeit~~ **Aufmerksamkeit!**

Referenzen

- Durrell, Martin, Paul Bennett, Silke Scheible, und Richard J. Whitt. 2012. *The GerManC Corpus*. University of Manchester. <http://hdl.handle.net/20.500.14106/2544>, https://www.ids-mannheim.de/fileadmin/lexik/uww/dateien/GerManC_Documentation.pdf (zugegriffen: 22.07.2024).
- Ehrmanntraut, Anton. 2024. „Historical German Text Normalization Using Type- and Token-Based Language Modeling“. ArXiv Pre-print. <https://doi.org/10.48550/arXiv.2409.02841>.
- Jurish, Bryan. 2012. *Finite-State Canonicalization Techniques for Historical German*. Dissertation, Universität Potsdam. <https://nbn-resolving.org/urn:nbn:de:kobv:517-opus-55789>.
- Jurish, Bryan, Marko Drotschmann, und Henriette Ast. 2013. „Constructing a Canonicalized Corpus of Historical German by Text Alignment“. In *New Methods in Historical Corpora*, 221–34. Tübingen: Narr.
- Krasselt, Julia, Marcel Bollmann, Stefanie Dipper, und Florian Petran. 2015. *Guidelines für die Normalisierung historischer deutscher Texte / Guidelines for Normalizing Historical German Texts*. Bochumer Linguistische Arbeitsberichte 15. <https://www.linguistics.rub.de/forschung/arbeitsberichte/15.pdf> (zugegriffen: 22.07.2024).
- Ljubešić, Nikola, Katja Zupan, Darja Fišer, und Tomaž Erjavec. 2016. „Normalising Slovene data: historical texts vs. user-generated content“. In *Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016*, hg. von Stefanie Dipper, Friedrich Neubarth, und Heike Zinsmeister. Bochumer Linguistische Arbeitsberichte 16. Bochum. https://www.linguistics.rub.de/konvens16/pub/19_konvensproc.pdf (zugegriffen: 22.07.2024).
- Odebrecht, Carolin, Laura Perlitz, Gohar Schnelle, und Catharina Fischer. 2020. *Annotationsrichtlinien zu Ridges Herbolgy Version 9.0*. Humboldt-Universität zu Berlin. https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/ridges-projekt/download-files/pubs/ridgesv9_2020-03.pdf (zugegriffen: 22.07.2024).
- Pettersson, Eva, Beáta Megyesi, und Jörg Tiedemann. 2013. „An SMT Approach to Automatic Annotation of Historical Text“. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA 2013*, hg. von Þórhallur Eypórsson, Lars Borin, Dag Haug, und Eiríkur Rögnvaldsson, 54–69. Oslo, Norwegen: Northern European Association for Language Technology. https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=87&Article_No=5 (zugegriffen: 22.07.2024).
- Scheible, Silke, Richard J. Whitt, Martin Durrell, und Paul Bennett. 2011. „A Gold Standard Corpus of Early Modern German“. In *Proceedings of the 5th Linguistic Annotation Workshop*, hg. von Nancy Ide, Adam Meyers, Sameer Pradhan, und Katrin Tomanek, 124–28. Portland, Oregon, USA: Association for Computational Linguistics. <https://aclanthology.org/W11-0415>.
- Xue, Linting, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, und Colin Raffel. 2021. „ByT5: Towards a token-free future with pre-trained byte-to-byte models“. ArXiv Pre-print. <https://doi.org/10.48550/arXiv.2105.13626>.

Appendix: Wie Zusammen-/Getrennt-Schreibung umsetzen

Zum erstenmal nahm er irgend ein Geräusch wahr .



Zum ersten_Mal nahm er irgend ein Geräusch wahr .



Zum ersten Mal nahm er irgendein Geräusch wahr .

Appendix: Wie viele Trainingsdaten sind nötig?

