

GSWNORM2022 - Shared Task on Text Normalization for Swiss German

Pius von Däniken^{1,*}, Manuela Hürlimann¹ and Mark Cieliebak²

¹Centre for Artificial Intelligence, Zurich University of Applied Sciences (ZHAW), Winterthur, Switzerland

Abstract

Written Swiss German is not standardized and varies across authors and their dialects and its use is almost exclusively constrained to communication on social media or via text messaging. Many corpora will therefore contain many distinct surface forms for the same word which can make their analysis challenging. It is therefore desirable to be able to normalize them to a single common surface form. We collected Swiss German utterances from social media and two annotators mapped every token to a corresponding form in Standard German. The task is to build models that can perform such a mapping automatically. This is different from translation since the resulting normalized utterance will in general not be grammatically correct Standard German as word order is preserved. A similar effort has previously been undertaken for text messages by the SMS4Science project. There is also a recent related shared task on lexical normalization of other languages at WNUT2021 workshop. During the shared task session, we presented the shared task dataset, how it was created, and gave an overview of the annotation tool. Since there were no participants this year, we presented the results of a naive baseline on the task dataset.


SwissText 2022: Swiss Text Analytics Conference, June 08–10, 2022,
Lugano, Switzerland

*Corresponding author.

✉ vode@zhaw.ch (P. v. Däniken); hueu@zhaw.ch (M. Hürlimann);
ciel@zhaw.ch (M. Cieliebak)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License
Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)