Cross-Lingual Sentiment Analysis

Generating datasets for German by leveraging MT

Jule Godbersen

23377840

Motivation In Sentiment Analysis (SA) the goal is to automatically determine the polarity of a given text. As "the majority of current sentiment analysis systems address a single language, usually English" (Dashtipour et al., 2016; p. 757), this can be seen as motivation to explore how SA systems can be improved for low-resource languages such as German. For example Waltinger (2010) make use of semi-automatic translation and thus were able to construct resources for German SA. More recently, Barriere and Balahur (2020) make use of a multilingual transformer model and leverage automatic translation to adapt the model to non-English languages, like German.

Research Questions The main objective of this project is to explore the influence of the training dataset on the performance of SA on a German test dataset. I will investigate the following two questions:

- How does the training dataset size influence the performance of SA in German?
- Is leveraging Machine Translation (MT) beneficial for SA in low-resource scenarios like German?

Data This project makes use of the Massively Multilingual Corpus of Sentiment Datasets (MMS) by Augustyniak et al. (2023). It contains text samples in English as well as in German, annotated as being "positive", "neutral", or "negative" each. Within this project the data will be preprocessed by e.g. balancing for sentiment labels. The translations will be done using the Google Translate via the deep-translator package.

Methodology Following Toledo-Ronen et al. (2020), I want to investigate a "translate-test" and a "translate-train" approach, but for the task of SA. In the appendix the specific scenarios are listed. The underlying multilingual model for the experiments will be "xlm-roberta-base" (Conneau et al., 2020). Its parameters will be frozen, while an adapter and a linear classification layer will be trained to finetune the model for the SA task.

References

Łukasz Augustyniak, Szymon Woźniak, Marcin Gruza, Piotr Gramacki, Krzysztof Rajda, Mikołaj Morzy, and Tomasz Kajdanowicz. Massively multilingual corpus of sentiment datasets and multi-faceted sentiment classification benchmark, 2023. URL https://arxiv.org/abs/2306.07902.

Valentin Barriere and Alexandra Balahur. Improving sentiment analysis over non-English tweets using multilingual transformers and automatic translation for data-augmentation. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 266–271, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.23. URL https://aclanthology.org/2020.coling-main.23.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8:757–771, 2016. URL https://link.springer.com/article/10.1007/s12559-016-9415-7.

Orith Toledo-Ronen, Matan Orbach, Yonatan Bilu, Artem Spector, and Noam Slonim. Multilingual argument mining: Datasets and analysis. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 303–317, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.29. URL https://aclanthology.org/2020.findings-emnlp.29.

Ulli Waltinger. GermanPolarityClues: A lexical resource for German sentiment analysis. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias, editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2010/pdf/91_Paper.pdf.

Appendix

As of now I want to implement the following scenarios to investigate my research questions. Note that DE represents data samples in German, and EN represents data samples in English.

Scenario	Training on	Testing on
Translate-Test	EN	DE translated into EN
Translate-Train	EN translated into DE	DE
Low-Resource	DE	DE
All-Resource	DE + EN translated into DE	DE

Table 1: Overview over scenarios