Character-Based Convolutional Neural Networks for Authorship Attribution of Sockpuppet Accounts

Tim Trappen
Ruhr-Universität Bochum
tim.trappen@ruhr-uni-bochum.de

In this work, we apply the method of character-based convolutional neural networks (CNN) (Zhang et al. 2015, Ruder et al. 2016) to the task of linguistic authorship attribution and verification of so called 'sockpuppet' accounts on Wikipedia talk pages. We try to connect the main account of a user suspected of sockpuppetry (called the 'sockpuppetteer') to a subset of possible sockpuppets by first training a model to attribute authorship of posts to the sockpuppetteer as an open-set, binary classification task. We choose alphanumeric and special characters as input features for the CNN, constructing three different input feature sets, two for English and one for German. The corpora used for training and evaluation of the model are small, consisting of conversations among users from German Wikipedia talk pages. Our results for the attribution task of the sockpuppetteer are competitive, with the German input feature set reaching a F1-score of 0.84. Compared to related research by Solorio et al. (2013), who approach a similar task using support vector machines (SVM) with an English dataset, our CNN approach sees an increase of 0.12 in F1-score. In a second step, we train additional models to attribute authorship to the sockpuppetteer when a suspected sockpuppet has been removed from the training data, and subsequently use that model to verify authorship between sockpuppetteer and the removed sockpuppet in a closed-set, binary classification task. While the results for the attribution task stay relatively consistent, the verfication task sees the model struggle for five distinct sockpuppet accounts, indicating that their use of characters is very similar to that of the sockpuppetteer. We analyze the misclassifications of the verification task for their relative token frequencies, and find that special characters, as well as certain function words are shared in frequency between the sockpuppetteer and potential sockpuppets. We conclude that the method of character-based CNNs performs well in binary open-set authorship attribution and closed-set authorship verification tasks, even when the available data is severely limited.

References: • Ruder, S., P. Ghaffari & J.G. Breslin (2016). Character-Level and Multi-Channel Convolutional Neural Networks for Large-Scale Authorship Attribution. CoRR, arXiv:abs/1609.06686. • Solorio, T., R. Hasan & M. Mizan (2013). A Case Study of Sockpuppet Detection in Wikipedia. In *Proceedings of the Workshop on Language Analysis in Social Media*. Atlanta, Georgia: Association for Computational Linguistics, 59-68. • Zhang, X., J. Zhao & Y. LeCun (2015). Character-level convolutional networks for text classification. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama & R. Garnett (eds.), *Advances in Neural Information Processing Systems 28 (NIPS 2015)*. Montreal, Canada: Curran Associates, Inc.