UNIVERSITY OF MALTA

Abstract

Faculty of ICT

M.Sc. Human Language Science and Technology

by Xiaoyu Bai

We present a novel study on CEFR level prediction in writings by non-native speakers of English, using on the one hand clean data elicited in a language learning context, and on the other hand noisy, spontaneous data from social media. The elicited data were drawn from the freely accessible learner corpus *EF-Cambridge Open Language Database* and consist of level-matched short essays written as part of an online English course. The spontaneous data were gathered from the social media platforms Twitter and Reddit, where users' self-reported proficiency levels were used as distant labels. Our level classification experiments were run both *within* and *across* the two domains as well on *mixed-domain* data. They were mainly conducted using linear SVM and logistic regression, although we also briefly explored bidirectional LSTM and convolutional neural networks, particularly in a setting of multi-task learning. We find that distant supervision based on user self-reports is a viable option of automatically generating noisily labelled training data for learner level prediction from social media texts. Despite the noisy nature of both the data and the labels, level prediction within the social media domain proved to be feasible, with system performance clearly beating the majority class baseline. Classification across the two domains, however, was revealed to be unsuccessful.

Keywords: Learner level classification, NLP, social media, machine learning