Comparing the Performance of CNNs and Shallow Models for Language Identification

Andrea Ceolin - Università di Modena e Reggio Emilia (ceolin@unimore.it)

Eighth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) - 2021

RDI

- The task requires distinguishing standard Romanian from Moldavian tweets (Găman et al. 2020) training on a corpus of news articles (i.e., cross-domain classification).
- CNNs match the performance of SVM and NB models when **data augmentation** is performed.
- The best performance, F_1=0.76, was obtained by shuffling the sentences (!), cf. Wei and Zou (2019). This leads to an improvement of $\approx$0.05 over a CNN baseline.
- The best team (SUKI) reached a test performance of F_1=0.777.
- Our submission was not great (F_1=0.653), probably because of overfitting.

ULI

- The task requires classifying texts from 178 languages from the Wanca 2017 corpus (Jauhiainen et al. 2020), with a focus on 29 Uralic varieties. The classes are largely imbalanced.
- We have not been able to train a CNN on the task, because of the high number of labels.
- A SVM model was trained to separate the 29 target Uralic varieties from the other languages, and a MNB model was trained to distinguish them.
- Combining the predictions of different shallow classifiers, we obtained state-of-the-art performance on the first track (F_1=0.81). **Rare languages** needed to be dealt with independently, because they were frequently misclassified.

DLI

- The task requires classifying three South Dravidian languages using a dataset of 17,672 YouTube comments (Chakravarthi et al. 2020a,b) involving code-switching.
- Using a CNN and performing **balanced sampling** yielded to a system that reached an F_1=0.9, in line with the performance of shallow models.
- The best team (LAST) reached a test performance of F_1=0.93.
- Our model seemed to slightly suffer from a poor representation of the 'other-languages' class, but the accuracy on the three main classes was in line with the other submissions.