Abstract

Solving automatic language identification problem is necessary for many tools that work with multilingual data. It drastically simplifies data collection for linguistic research, especially when it comes to previously poorly researched languages. This thesis focuses on large-scale language identification for minority languages of Russia.

We collected data for 48 languages spoken on the territory of the Russian Federation. We experimented with different traditional approaches with Support Vector Machines (SVMs) and neural networks. We explored dataset size limitations and cross-domain testing, as well as experimented with balanced and unbalanced data. We achieved 94.15% accuracy for the *in-domain* testing with an SVM model using character n-grams. The best result for testing on an *out-of-domain* dataset (82.51%) was achieved with a 2-step SVM model, by predicting the language family first and then classifying the languages within each family separately. Our experiments within Deep Learning approaches did not yield as high results as approaches with SVMs, in line with the trend seen in previous works on language identification.