Authorship Detection with PPM   
Victoria Bobicev

*Technical University of Moldova*

*victoria\_bobicev@rol.md*

This paper reports on our work in the PAN 2013 author identification task. The task is to automatically detect the author of the given text having small training sets with known authors. We solved the authorship identification task as a type of classification task using similarity-based method. We applied PPM (Prediction by Partial Matching) compression algorithm based on character n-gram statistical model which is a back-off smoothing technique for finite-order Markov models. PPM produces a language model and can be used in a probabilistic text classifier. The PPM technique uses character context models to build an overall probability distribution for predicting upcoming characters in the text. PPM is a special case of the general blending strategy. The PPM models use an escape mechanism to combine the predictions of all character contexts of length m, where m is the maximum model order. Our similarity function for text classification was cross-entropy of the test document calculated using the probabilities estimated on the training corpus. The lower is entropy, the more similar are texts.

In this task we were given a small set of "known" documents by a single person and a "questioned" document and the task was to determine whether the questioned document was written by the same person who wrote the known document set. In this case we could not use standard classification technique as we had only one class. In order to make the decision we divided all given known and unknown texts in small fragments and compared their entropies calculated on the basis of known texts models and on the basis of unknown text models. If the entropies were considerably different we considered that texts were written by two different authors.

The preliminary results obtained on the testing data released by organizers are following: for the 10 English test texts 9 were detected correctly; for the 20 Greek test texts 13 were detected correctly; for the 5 Spanish test texts 3 were detected correctly.

The final results are the following: performance on all test data: F-measure = 0.655, Precision = 0.663, Recall = 0.647; performance on English data: F-measure = 0.644, Precision = 0.655, Recall = 0.633; performance on Greek data: F-measure = 0.712, Precision = 0.724, Recall = 0.700; performance on Spanish data: F-measure = 0.600, Precision = 0.600, Recall = 0.600.