

Domain-independent Classification of Automatic Speech Recognition Texts

Evgeniya Mescheryakova^{1,2} and Lyubov' Nesterenko^{1,2}

¹ DC-Systems,

² National Research University – Higher School of Economics

Abstract. Call centers receive large amounts of incoming calls. The calls are being regularly processed by the analytical system, which helps people automatically inspect all the data. Such system demands a classification module that can determine the topic of conversation for each call. Due to high costs of manual annotation, the input for this module is the automatically transcribed calls. Hence, the texts (=automatic transcription) used for classification contain ill-transcribed words which can probably influence the classification process. Another important point is that this module also has special requirements: it should be domain-independent and easy to setup. Document classification task always requires an annotated data set for classifier training, but it seems to be too costly to make an annotated training set for each domain manually. In this paper, we propose an approach to automatic speech recognition texts classification that allows the user avoiding full manual annotation and at the same time to control its quality.

Keywords: document classification, document clustering, automatic speech recognition, noisy texts processing