Contribution ID: **302** Type: **Sectional reports**

Building corpora of transcribed speech from open access sources

Thursday 13 September 2018 14:45 (15 minutes)

Currently there are hardly any open access corpora of transcribed speech in Russian that can be effectively used to train those speech recognition systems that are based on deep neural networks—e.g., DeepSpeech. This paper examines the methods to automatically build massive corpora of transcribed speech from open access sources in the internet, such as radio transcripts and subtitles to video clips.
Our study is focused on a method to build a speech corpus using the materials extracted from the YouTube video hosting. YouTube provides two types of subtitles: those uploaded by a video's author and those obtained through automatic recognition by speech recognition algorithms. Both have their specifics: author subtitles may have timing inaccuracies, while automatically recognized subtitles may have recognition errors.
We used the YouTube Search API to obtain the links to various Russian-language video clips with subtitles available—words from a Russian dictionary served as an input. We examined two strategies to extract audio recordings with transcripts corresponding to them: by using both types of subtitles or only those that were produced through automatic recognition. The voice activity detector algorithm was employed to automatically separate the segments.
Our study resulted in creating transcribed speech corpora in Russian containing 1000 hours of audio recordings. We also assessed the quality of obtained data by using a part of it to train a Russian-language automatic speech recognition system based on DeepSpeech architecture. Upon training, the system was tested on a data set consisting of audio recordings of Russian literature available on voxforge.com—the best WER demonstrated by the system was 18%.

Authors: SHALEVA, Anna (Saint-Petersburg State University); FEDOSEEV, G. (Saint-Petersburg State University); IAKUSHKIN, Oleg (Saint-Petersburg State University)

Presenter: SHALEVA, Anna (Saint-Petersburg State University)

Session Classification: 11. Big data Analytics, Machine learning

Track Classification: 11. Big data Analytics, Machine learning