Building Speech Corpus Resources Using NooJ

Lesia Kaigorodova
United Institute of Informatics Problems of the NAS of Belarus, Minsk, BELARUS
Lesia.Piatrouskaya@gmail.com

Yuras Hetsevich
United Institute of Informatics Problems of the NAS of Belarus, Minsk, BELARUS
Yuras.Hetsevich@gmail.com

Abstract

Creating a speech corpus is usually a hard and time-consuming process. There are no free or cheap available datasets that anyone could use for real-life speech recognition tasks. There are several reasons for this. Firstly, very often it involves lots of manual labelling. Secondly, available data are usually unstructured datasets from different resources or have a format that does not fit real needs. These datasets should be preprocessed in some way in order to get the unified format that would be further used by speech recognition systems.

In this project we create speech corpus using the movie database with the subtitles in Belarusian and NooJ module.

First, we use NooJ Syntactic Grammar in order to parse text subtitles files. Here we extract information and add such annotations as start time of the speech, end time of the speech from the movie track and text information of the speech itself. So far we would have a labeled dataset *{ audio signal, text }, where audio signal* represents features and *text* represents labels. This is the standard format of the data that is widely used for training and testing machine learning models for speech recognition tasks.

Second, in this project we use NooJ module to extend feature space of our dataset. We decided to add as much features as possible and then filter them out while training and tuning our machine learning model. We use annotations from existing Belarusian resources (dictionaries, morphological and syntactical grammars). Thus we add not only such kind of information as morphology and syntax, but also transcription and prosody that are known to be highly valuable for speech recognition systems. As a drawback, we also do understand that this method may also involve some minor correction with manual labelling, but it is not as critical as building such features from scratch.

We could see that the approach for creating speech recognition corpus using NooJ tool may be time saving and allows us to build more intelligent engines for speech-to-text systems.

Referencies

1. Silberztein Max, 2018. NooJ Manual. Available for download at: http://www.nooj-association.org. Date of access : 30.01.2019.
2. Hetsevich, Y. Semi-automatic Part-of-Speech Annotating for Belarusian Dictionaries Enrichment in NooJ / Y. Hetsevich, V. Varanovich, E. Kachan, I. Reentovich, S. Lysy // Automatic Processing of Natural-Language Electronic Texts with NooJ: 10th International Conference, NooJ 2016, České Budějovice, Czech Republic, June 9-11, 2016, Revised Selected Papers / ed. L. Barone, M. Monteleone, M. Silberztein. — Springer, 2017. — P. 101-111.
3. Lysy, S. Addition of IPA Transcription to the Belarusian NooJ Module / S. Lysy, H. Stanislavenka, Y. Hetsevich // Automatic Processing of Natural-Language Electronic Texts with NooJ: 10th International Conference, NooJ 2016, České Budějovice, Czech Republic, June 9-11, 2016, Revised Selected Papers / ed. L. Barone, M. Monteleone, M. Silberztein. — Springer, 2017. — P. 14-22.
4. Zahariev, V. Grapheme-to-Phoneme and Phoneme-to-Grapheme Conversion in Belarusian with NooJ for TTS and STT Systems / Vadim Zahariev, Stanislau Lysy, Alena Hiuntar, Yury Hetsevich // Automatic Processing of Natural-Language Electronic Texts with NooJ: 9th International Conference, NooJ 2015, Minsk, Belarus, June 11-13, 2015, Revised Selected Papers / ed. T. Okrut, Y. Hetsevich, M. Silberztein, H. Stanislavenka. — Springer International Publishing, 2016. — P. 137-150.
5. Hetsevich, Yu. Grammars for Sentence into Phrase Segmentation: Punctuation Level / Yuras Hetsevich, Tatsiana Okrut, Boris Lobanov // Automatic Processing of Natural-Language Electronic Texts with NooJ: 9th International Conference, NooJ 2015, Minsk, Belarus, June 11-13, 2015, Revised Selected Papers / ed. T. Okrut, Y. Hetsevich, M. Silberztein, H. Stanislavenka. — Springer International Publishing, 2016. — P. 74-82.