Title: Normalization of numbers into spoken form for text-to-speech systems

Author: Jakub Růžička

Institute: Institute of Formal and Applied Linguistics

Supervisor: Mgr. et Mgr. Ondřej Dušek, Ph.D., Institute of Formal and Applied Linguistics

Abstract: A necessary part of any text-to-speech system is the normalization of numbers and words containing numbers. The accuracy of this process can significantly affect the quality of the resulting speech. The main goal of this work is the design and implementation of a number normalization module for Czech. Words containing digits are first assigned to one of the predefined categories. Based on the category given, possible spoken forms are subsequently generated. For the selection of the contextually correct variant, an existing language model is used. The system is distributed as a Python package and can run on Linux or in a Docker container whose configuration is part of the project. Moreover, a specialized data annotation application has been designed and written for creating the datasets for the Czech text normalization task. Two datasets with 1,882 sentences and 3,185 words requiring normalization were obtained using the data annotation service. The system achieved a sentence-level accuracy of over 80% on both datasets. We perform a detailed error analysis based on the results, and propose further improvements.

Keywords: Czech text normalization, number normalization, text-to-speech systems, weighted finite-state transducer