

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра дискретной математики и алгоритмики

Аннотация к дипломной работе

«Разработка системы распознавания белорусской речи»

Тылецкий Арсений Витальевич

Научный руководитель – старший преподаватель кафедры дискретной математики и алгоритмики ФПМИ Пирштук Д. И.

Минск, 2024

Реферат

Дипломная работа, 53 страницы, 18 рисунков, 8 таблиц, 18 источников

Ключевые слова: РАСПОЗНАВАНИЕ РЕЧИ, ЗВУК, ПРЕОБРАЗОВАНИЕ ЗВУКА, WER, LIBRISPEECH, COMMON VOICE, FLEURS, HMM, GMM, CTC, RNN-T, МЕХАНИЗМ ВНИМАНИЯ, ТРАНСФОРМЕР, НЕЙРОННЫЕ СЕТИ, WHISPER.

Объектом исследования является задача распознавания белорусской речи.

Предметом исследования являются модели нейронных сетей для распознавания речи, основанных на механизме внимания.

Целью работы является улучшения качества распознавания белорусской речи.

В ходе работы были рассмотрены постановка задачи распознавания речи, методика оценки качества решения данной задачи, ее применение. Был произведен обзор наборов данных для распознавания речи, а также более подробно рассмотрены мультязычные датасеты Common Voice и FLEURS, содержащие примеры распознавания белорусской речи. Сделан обзор и сравнение различных методов решения задачи распознавания речи, как исторических, основанных на моделях со скрытыми марковскими цепями (HMM), так и более современных, таких как Connectionist temporal classification (CTC), Recurrent Neural Network Transducer (RNN-T) и моделей, основанных на механизме внимания. Был рассмотрен один из самых современных end-to-end алгоритмов для распознавания мультязычной речи Whisper, а также показано неудовлетворительное качество распознавания белорусской речи с помощью этого алгоритма. Произведя предварительную обработку аудио и текста, мы обучили несколько моделей Whisper размеров Tiny и Small на датасете Common Voice 16.1, значительно улучшив качество распознавания белорусской речи. Мы подробно исследовали основные ошибки распознавания полученной модели, предложив и проанализировав различные пути решения выявленных проблем. Дополнительно были исследованы качество распознавания знаков препинания и забывание моделью русского языка.

Полученная в результате работы модель имеет хорошее качество распознавания белорусской речи, ошибаясь в среднем лишь в каждом 7 слове, в то время как лучшая оригинальная модель Whisper ошибается в каждом 2-3 слове.

Abstract

Diploma thesis, 53 pages, 18 figures, 8 tables, 18 sources.

Keywords: SPEECH RECOGNITION, SOUND, SOUND TRANSFORMATION, WER, LIBRISPEECH, COMMON VOICE, FLEURS, HMM, GMM, CTC, RNN-T, ATTENTION MECHANISM, TRANSFORMER, NEURAL NETWORKS, WHISPER.

The object of research is the task of Belarussian speech recognition.

The subject of study is models of neural networks for speech recognition, based on attention mechanism.

The aim of this work is to enhance quality of Belarussian speech recognition.

In the course of the work we considered the formulation of the speech recognition problem, the methodology for estimating the quality of problem solution and its application. A review of the speech recognition datasets was carried out, as well as the multilingual Common Voice and FLEURS datasets containing examples of Belarussian speech recognition were considered in more detail. We provided a review and comparison of various methods for solving the speech recognition problem, both historical, based on models with hidden Markov chains (HMM), and more modern, such as Connectionist temporal classification (CTC), Recurrent Neural Network Transducer (RNN-T) and models based on the attention mechanism. We also considered Whisper – one of the most modern end-to-end algorithms for multilingual speech recognition and showed the unsatisfactory quality of Belarussian speech recognition using this algorithm. After preprocessing audio and text, we trained Tiny and Small Whisper models on the Common Voice 16.1 dataset, significantly improving the quality of Belarussian speech recognition. We have investigated in detail the main recognition errors of the resulting models, suggesting and analyzing various ways to solve the identified problems. Additionally, the quality of punctuation recognition and the model's forgetting of the Russian language were investigated.

The resulting model has good quality of Belarussian speech recognition, making mistakes on average only in every 7th words, while the best original Whisper model makes mistakes in every 2nd-3rd word.