УДК 004.522

АНАЛИЗ РЕЧЕВЫХ КОРПУСОВ ДЛЯ ОБУЧЕНИЯ СИСТЕМ РАСПОЗНАВАНИЯ РУССКОЙ РЕЧИ

Капуста К.Л.
(Университет ИТМО, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук)
Научный руководитель – д.т.н. Карпов А.А
(Университет ИТМО, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук)

Введение. Распознавание речи является важным направлением в области искусственного интеллекта и имеет широкое применение в таких областях, как медицина, телекоммуникации, автоматизация производственных процессов и многие другие. Для достижения высокой точности распознавания речи, помимо грамотной работы с моделями, необходимо использовать качественные корпуса для обучения систем распознавания речи. В данной работе проведен анализ различных открытых корпусов аннотированной речи на русском языке с целью выявления наиболее подходящих наборов данных для обучения современных моделей.

Основная часть. Для обучения систем распознавания речи используются речевые корпуса, которые представляют собой набор аудиофайлов с транскрипциями. Аудиофайлы содержат звуковой сигнал, который может отличаться качеством (отношение сигнал/шум). Сама речь диктора может содержать паузы хезитации, речевые ошибки, различный диалектный или социолектный оттенок и другие особенности. Отличаться могут в том числе и транскрипции. По способу создания они могут быть полностью ручными, автоматическими (полученными при помощи других систем распознавания речи) или полученными при помощи принудительного выравнивания. Самой качественной транскрипцией обычно является ручная. Для достижения высокой точности распознавания речи необходимо использовать качественные корпуса, которые содержат большое количество разнообразной, но при этом распознаваемой на слух речи. Проблемными особенностями могут выступать ошибки в транскрипции, наличие сильного фонового шума и наличие нецензурной лексики в случае использования языковой модели, которая её в себя не включает. Одним из самых крупных открытых корпусов русской речи является Open STT. Он содержит около 20000 часов речи и 16 миллионов высказываний различного происхождения [1]. Open STT по сути является набором из нескольких корпусов с различными видами аннотации, качеством и происхождением записей. Он включает в себя корпуса речи с радиостанций, аудиокниг, видеозаписей, звонков и прочих источников [2]. В рамках данного исследования также были выбраны открытые корпуса Golos и Multilingual TEDx из OpenSLR. Golos состоит из 1240 часов русской речи с ручной аннотацией [3]. Multilingual TEDx состоит из набора выступления TED Talks на разных языках с ручной аннотацией, объём русскоязычных данных составляет около 62 часов речи [4]. Цель работы заключается в анализе данных речевых корпусов и выявлении наиболее адекватных из них для обучения современной нейросетевой модели распознавания речи.

Выводы. В результате работы проведен сравнительный анализ свободнодоступных корпусов русской речи по основным характеристикам, таким, как объём данных, число дикторов, качество аннотации, качество речевого и звукового материалов и другим. Результаты исследования могут помочь составить наиболее подходящую выборку для обучения системы автоматического распознавания речи на русском языке и улучшения качества работы существующих систем распознавания речи.

Список использованных источников:

1. Slizhikova, A., Veysov, A., Nurtdinova, D., Voronin, D., Baburov, Y. Russian open speech to text (stt/asr) dataset v1.0 [Электронный ресурс]. – URL: https://github.com/snakers4/open_stt/ (дата обращения: 15.02.2023).
2. Andrusenko A., Laptev A., Medennikov I. Exploration of end-to-end asr for openstt–russian open speech-to-text dataset // Speech and Computer: 22nd International Conference, SPECOM 2020, St.

Petersburg, Russia, October 7–9, 2020, Proceedings 22. – Springer International Publishing, 2020. – С. 35-44.

3. Karpov N., Denisenko A., Minkin F. Golos: Russian dataset for speech research // arXiv preprint arXiv:2106.10161. – 2021.

4. Salesky E. et al. The multilingual tedx corpus for speech recognition and translation //arXiv preprint arXiv:2102.01757. – 2021.