УДК 004.93

РАЗРАБОТКА СИСТЕМЫ РАСПОЗНАВАНИЯ БЕЛОРУССКОЙ РЕЧИ

А. В. Тылецкий

Белорусский государственный университет,
пр. Независимости, 4, 220030, г. Минск, Беларусь, a.tyletsky@gmail.com.
Научный руководитель: Д. И. Пирштук, старший преподаватель

Рассмотрены постановка задачи автоматического распознавания речи, основная метрика оценки ее качества, задача мультиязычного распознавания речи. Для решения задачи распознавания белорусской речи были дообучены 2 модели серии Whisper размера Small и Tiny. Данные модели показывают значительное улучшение качества распознавания белорусской речи на внешнем тестовом датасете. В результате анализа основных ошибок модели предложен способ дальнейшего улучшения качества распознавания.

Ключевые слова: распознавание белорусской речи; WER; Whisper; FLEURS; Common Voice.

Введение

Одной из проблем построения качественной системы распознавания речи является необходимость обучения глубоких нейронных сетей на большом объёме данных. Это уже не является проблемой для, например, таких популярных языков как английский, однако может являться проблемой для таких языков как белорусский. В последние несколько лет стали показывать относительно хорошее качество модели для распознавания мультиязычной речи. Данные модели могут быть дополнительно дообучены на задачу распознавания речи определенного непопулярного языка, например, белорусского.

Общая постановка задачи автоматического распознавания речи

Пусть у нас есть некоторый аудиофайл X, на котором записана человеческая речь. Наша задача представить человеческую речь в виде последовательности меток $L = (l_1, l_2, \ldots, l_n)$, где каждая метка $l_i \in V$, где V – некоторый словарь меток. Метки различаются в зависимости от языка говорящего и обычно являются словами, но могут быть и, например, иероглифами или отдельными буквами. Через V^* обозначим набор всевозможных последовательностей меток из словаря V. Тогда более формально задача автоматического распознавания речи по данному аудиофайлу X найти $L^* \in V^*$ такой, что:

$$L^* = \arg\max_{L \in V^*} p(L|X) \quad (1)$$

Звук может кодироваться различным образом, однако наиболее привычный способ представления звука есть пара (r, A), где $r \in \mathrm{N}$ – частота дискретизации сигнала, $A \in [-1;1]^n$ – массив амплитуд сигнала во времени.

Даже очень маленькие аудиофайлы представляют собой огромные последовательности чисел, которые нельзя сразу же подать на вход модели машинного обучения. Во-первых, это делать достаточно неэффективно и модель будет долго работать, прежде чем выдать результат. Во-вторых, модели сложно извлечь данные из обычной последовательности чисел. Поэтому перед применением модели машинного обучения звук нужно обработать и извлечь полезные признаки. Наиболее популярным является метод получения мелкепстральных коэффициентов. Здесь и далее можно считать, что под аудиофайлом X подразумевается матрица признаков, извлеченных из звука.

Метрики точности решения задачи распознавания речи

Наиболее общепринятой метрикой решения задачи распознавания речи является метрика Word Error Rate (сокращенно WER). Она высчитывается как нормализованное на длину целевой последовательности слов минимальное количество операций замены слова, вставки слова, удаления слова из предсказанной последовательности слов, для того чтобы получить из нее целевую последовательность слов.

Метрика имеет ряд проблем. Например, она никак не взвешивает ошибки. Очевидно, что если правильное слово отличается одной буквой от предсказанного слова, то такая ошибка может быть не так критична, как пропуск целой частицы «не», которая может полностью поменять смысл предложения.

Распознавание мультиязычной речи

Качественное решение задачи распознавания мультиязычной речи с помощью одной модели долго не была решена. Одними из первых хороших моделей в данной области стали модели серии Whisper [1]. Whisper модели обучались на 680000 часов аудио, из которых 65% данных есть примеры распознавания английской речи, 17% – примеры распознавания речи других языков, а оставшиеся 18% – примеры перевода других языков на английский язык. При этом даже самая большая модель Large-v3 показывает при распознавании белорусской речи на внешнем датасете FLEURS [2] метрику WER=42.5%, в то время как при распознавании русской речи модель показывает WER=5.0%. Это связано с недостатком обучаемых данных. Модель видела в 130 раз меньше данных для белорусского языка, чем для русского.

Распознавание белорусской речи

Наиболее крупным общедоступным набором белорусской речи является Common Voice [3]. Последняя версия 16.1 вышла 5 января 2024 года, и его белорусская часть содержит 1694 часа аудио в формате MP3, которые весят суммарно около 32 ГБ и состоят из более чем 8000 различных голосов. Всего получается около 400000 примеров. Эти данные мы будем использовать для дополнительного дообучения Whisper моделей на задачу распознавания белорусской речи.

Мы обучили 2 версии модели Whisper: Tiny, как самую быструю, но не точную, и Small, как достаточно точную и быструю модель, при этом все еще доступную в обучении на стандартной видеокарточке Tesla P100-16GB. Обучение происходило с размером батча 32 в течении 4500 шагов для версии Tiny и в течении 6500 шагов для версии Small. Использовалась стандартная кросс-энтропийная функция потерь. В качестве оптимизатора был выбран Adam со стандартными параметрами. Скорость шага градиентного спуска в обучении изменяется циклически: увеличивается и уменьшается линейно между 0 и максимальным значением 0.0001. Каждые 500 шагов делалась валидация и сохранялась лучшая модель. Модель стабильно обучалась полностью в FP16 режиме. Это значит, что во время обучения все веса, активации, градиенты хранятся и вычисляются в точности FP16.

На валидации замерялись следующие метрики: WER и нормализованный WER (WER на нормализованной транскрипции), однако лучшая модель выбиралась именно по нормализованному WER. Обучение при этом будет происходить на сырых транскрипциях. Таким образом, наша модель будет сразу учиться расставлять заглавные буквы и знаки препинания, но при этом мы будем стремиться получить модель, которая лучше распознает сами слова, а не лучше расставляет знаки препинания в ущерб распознаванию слов.

Приведем замер качества обученных моделей на тестовой части выборки датасета FLEURS, содержащей белорусскую речь. Датасет FLEURS содержит сразу нормализованные транскрипции, поэтому замерять обычную метрику WER нет смысла. В следующей таблице собраны среднее и медианное значения WER на нормализованном предсказании модели для различных моделей: оригинальных моделей Tiny, Small, Large-v3, а также на дообученных на

белорусском сегменте данных моделей размера Tiny и Small. Также дана дополнительная информация про количество параметров и относительную скорость моделей.

Значения WER на тестовой белорусской части FLEURS

Модель	Норм. WER (среднее)	Норм. WER (медиана)	Количество параметров	Относительная скорость
Tiny	99,8	95,0	39M	32
Small	80,0	77,3	244M	6
Large-v3	43,8	42,4	1550M	1
Tiny (дообучение)	34,0	31,2	39M	32
Small (дообучение)	21,2	14,7	244M	6

Проведя анализ крупных ошибок модели (WER>100%) можно сделать вывод, что модели сложно распознавать речь с иностранными или редко встречающимися в белорусской речи словами. Следствием этого является то, что модель начинает многократно выдавать одну и ту же букву, слово или даже целую фразу. Тем не менее, большинство ошибок модели, особенно дообученной Small версии, являются незначительными. Например, модель может не поставить букву в конце слова, может наоборот поставить лишнюю букву в слове, а может перепутать падеж слова.

Решить вышеописанные проблемы может использование дополнительной постобрабатывающей распознанный текст модели. Такой способ является медленным, однако качественным. Данная модель должна принимать на вход текст и выдавать на выход также текст, со всеми исправленными ошибками. Собрать данные для такой модели достаточно просто: можно брать в качестве целевого текста текст из транскрипции, а в качестве входного текста – текст, распознанный текущей моделью. Примечательно, что данный способ также может расставлять недостающие знаки препинания или наоборот убирать лишние знаки.

Библиографические ссылки

1. Robust speech recognition via large-scale weak supervision / A. Radford [et al.] // International Conference on Machine Learning. 2023. P. 28492-28518.

2. FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech / A. Conneau [et al.] // 2022 IEEE Spoken Language Technology Workshop (SLT). 2023. P. 798-805.

3. MLS: A Large-Scale Multilingual Dataset for Speech Research / V. Pratap [et al.] // Interspeech. 2020.