УДК 81’322.5

ПИСАТЕЛЬ РОБИН ДРАНАТТАГОР: АПРОБАЦИЯ МОДЕЛИ WHISPER НА РУССКОЯЗЫЧНОЙ ЗВУЧАЩЕЙ РЕЧИ

Евгения Олеговна Колпащикова
стажер-исследователь лаборатории языковой конвергенции
Национальный исследовательский университет
«Высшая школа экономики» – Санкт-Петербург
190068, Санкт-Петербург, наб. канала Грибоедова, 123. eokolpaschikova@edu.hse.ru

Whisper – модель автоматического распознавания речи, презентованная компанией OpenAI в сентябре 2022 г. Whisper был обучен на 680 тыс. часах многоязычной и многозадачной речи, что должно было улучшить качество распознавания акцентов и сделать модель менее чувствительной к шуму на фоне разговора. Проведенное исследование посвящено апробации возможностей Whisper на полевых аудиозаписях из корпуса «Один речевой день» и оценке точности автоматически полученных расшифровок по сравнению с расшифровками, выполненными экспертами вручную. В статье описаны и сгруппированы допущенные моделью ошибки и выявлены другие особенности ее работы: к примеру, Whisper достаточно точно восстанавливает «проглоченные» говорящим слоги, однако не всегда верно улавливает грамматические окончания.

Ключевые слова: автоматическое распознавание речи; русский язык; повседневная речевая коммуникация; автоматическое и экспертное транскрибирование; Whisper.

Введение

Whisper представляет собой мультимодальную генеративную модель, выпущенную компанией OpenAI в сентябре 2022 г. Название было выбрано из-за близости слова *whisper* ‘шепот/шептать’ к аббревиатуре, которой создатели модели обозначили свой подход – WSPSR (Web-scale Supervised Pretraining for Speech Recognition – контролируемое предварительное обучение модели для распознавания речи в сетевом масштабе) [Radford et al. 2022]. Whisper использует звучащую речь в качестве входных данных и генерирует расшифровку на ее основании [Lin 2023].

Согласно создателям модели, ее целью ставилось изучение возможностей систем обработки речи, обученных предсказывать большие объемы расшифровок аудио в Интернете [Radford et al. 2022]. Модель была обучена на 680 тыс. часах многоязычного и многозадачного наблюдения. В качестве обучающего материала использовались аудиозаписи из Интернета, для которых существовала проверенная расшифровка [там же]. Благодаря разнообразию этого материала, Whisper должен уметь справляться с расшифровкой вне зависимости от скорости речи и акцента говорящего [Lin 2023].

Дополнительным функционалом Whisper является проставление знаков препинания в транскриптах [Документация Whisper]. В настоящей работе мы не обращались к этому аспекту расшифровок, однако такие возможности модели могут послужить основой для дальнейших исследований. Например, Л. Грис и его соавторы, одними из первых исследовавшие эффективность Whisper, оценивали работу модели на материалах виртуального португальского музея личных историй именно с точки зрения правильной (по сравнению с экспертной расшифровкой) расстановки точек, вопросительных и восклицательных знаков, а также двоеточий и даже точек с запятой [Gris et al. 2023].

Материал и методы исследования

Материалом для апробации Whisper в данном исследовании стали звукозаписи из корпуса ОРД. Его особенность и преимущество в том, что он содержит звукозаписи повседневного общения [Asinovsky et al. 2009]. Основной целью корпуса является «изучение устной речи, бытовой и профессиональной коммуникации» [Богданова-Бегларян et al. 2019]. В ходе работы над корпусом был получен представительный объем звукозаписей речевого материала более чем от 130 информантов [там же]. Это самая крупная в России коллекция звукозаписей повседневного общения, которая предназначена для проведения научных исследований в области фонетики, лек-

сики, морфологии, синтаксиса и прагматики устной речи, а также для поддержки прикладных разработок в области речевых технологий. До настоящего времени расшифровка звукозаписей (их перевод из звука в текст) велась экспертами вручную, что представляет собой весьма трудоемкий процесс. По этой причине не все звукозаписи корпуса до сих пор переведены в текст и задача получения их автоматических расшифровок является по-настоящему актуальной.

Появление высокоэффективных моделей распознавания речи, подобных Whisper, открывает перспективные возможности для создателей корпусов устной речи в плане наращивания объемов расшифрованного материала. Данная работа посвящена тестированию Whisper на полевых звукозаписях корпуса ОРД.

В ходе исследования с помощью Whisper был расшифровано около 200 речевых эпизодов, именуемых макроэпизодами [Sherstinova 2015]) из корпуса «Один речевой день», для каждого из которых существует экспертная разметка, выполненная вручную специалистом.

Для оценки качества работы моделей автоматической расшифровки применяется ряд параметров. Одним из наиболее часто используемых является Word Error Rate (WER). Этот метод основывается на расчете расстояния Левенштейна, или редакционном расстоянии. Редакционное расстояние представляет собой минимальное количество добавлений, удалений и замен слов, которые необходимо осуществить для получения одной строки из другой [Левенштейн 1965].

Для подсчета WER словесная последовательность, сформированная системой распознавания, сравнивается с референтной словесной последовательностью [Ali, Renals 2018]. Затем подсчитывается количество ошибок: сумма замененных слов, добавленных слов и удаленных слов; число ошибок, разделенное на общее число слов в последовательности, и называется показателем WER [там же]. Нормализация по общему числу слов необходима из-за того, что масштаб редакционного расстояния зависит от длины строки [McCowan et al. 2005]. Для подсчета WER с помощью языка программирования Python существует пакет JiWER, позволяющий быстро рассчитать WER и несколько других статистических параметров точности расшифровки (Match Error Rate, Word Information Lost и др.) [Документация JiWER].

К минусам WER относят необходимость экспертной разметки (некоего «золотого стандарта») для осуществления сравнения [Ali, Renals 2018]. Другими недостатками являются недостаточная глубина оценки, сложность в интерпретации и ограниченная гибкость с точки зрения назначения разного веса словам [McCowan et al. 2005].

Результаты и их оценка

Средний показатель WER для всех речевых эпизодов составил 49%. Минимальное значение оказалось равным 9%, максимальное – 97%.

На Рисунке представлено значение WER для каждого эпизодов (по горизонтали – номер эпизода, по вертикали – значение WER).

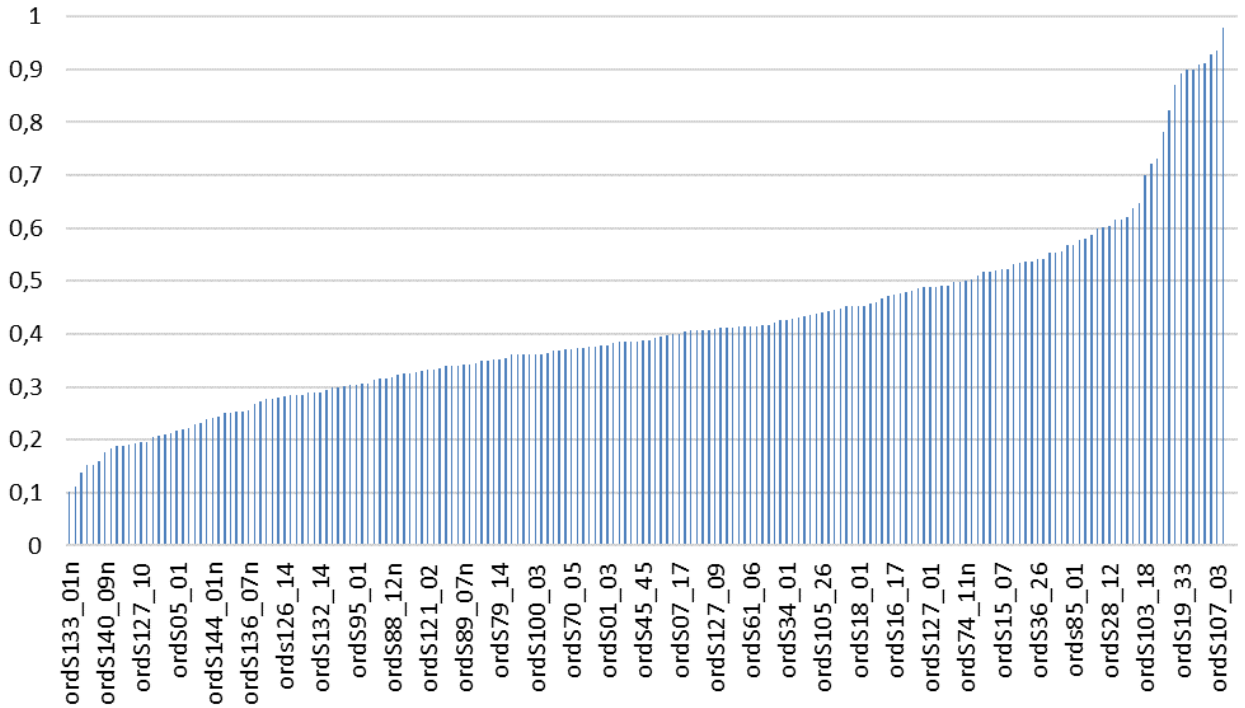

Рисунок. WER для макроэпизодов исследуемой выборки

Обратимся к сравнению расшифровок конкретных эпизодов для выявления особенностей работы модели Whisper.

Начнем с речевого эпизода ordS26-02. WER для его расшифровки в Whisper составил 16%. Главное преимущество этого речевого эпизода с точки зрения «удобства» распознавания заключается в том, что это размеренная монологическая речь почти в идеальной тишине. Говорящий заполняет анкету о своем восприятии устной речи, иногда проговаривая вопросы, но чаще просто комментируя свои ответы.

В этом речевом эпизоде говорящий часто «проглатывает» слоги, что вызвано скорее манерой речи, чем торопливостью. Во всех этих случаях модель Whisper правильно «достроила» то, что смогла расслышать, до полного слова. Также в этом эпизоде упоминаются несколько имен собственных. Whisper справилась с ними не идеально: *Ираклий Андроников* остался обозначенным как *Ирак и Андроников*, *Рабиндранат Тагор* получил новое имя – *Робин Дранаттагор*. Поскольку речь говорящего по большей части посвящена достаточно сложному предмету – коммуникации в эпоху быстрого развития компьютерных технологий, некоторые сложные слова, которые он использует, оказались затранскрибированы как два или несколько отдельных. Слово *эталонный* моделью Whisper было при обоих вхождениях распознано как *талонный*, а *речеслуховой аппарат* – как *речь о слуховой аппарат*. Иногда встречаются небольшие ошибки в глагольных и падежных окончаниях, например: *возникает ли <> трудности* (вместо *возникают*) или *речь представителями средств массовой информации* (вместо *представителей*).

Для эпизода ordS121-04 WER составил 15%. В этом речевом эпизоде происходит обсуждение стилистики оформления свадьбы; из контекста аудиозаписи можно предположить, что это запись курса, посвященного свадебной флористике. К причинам, по которым этот речевой эпизод был распознан достаточно хорошо, можно отнести формат «лекции» – почти не прерываемого монолога, а также отсутствие большого числа имен собственных. Вместе с тем лексика, используемая в рассказе о подборе украшений для свадьбы, не настолько специфична, чтобы модель ее не «знала». На протяжении всего речевого эпизода микрофон находится близко к говорящему, его очень хорошо слышно, а шумы на фоне лишь изредка перекрывают его голос. Whisper попытался распознать ответы на вопросы лектора, которые слышно достаточно плохо; расшифровать эти реплики получилось лишь отчасти: так, фраза *если рокеры <> закажут* превратилось в *други <> закажут*, *тонах* – в *то надо, я всех приглашу* – в *я всем предложу*. В некоторых специфических словах Whisper ошибается в орфографии (например, *сереневатые, цветы в УАЗе* вместо *в вазе*), но зато хорошо улавливает частицы: в речи говорящего часто встречаются *да* и *вот* и в расшифровке это передано. Большая часть ошибок приходится на окончания: *карий <> глаза*, *к прически*, *букета буду* (вместо *букет я*).

Для эпизода ordS130-08 показатель WER также равен 15%. Он представляет собой телефонный разговор на разные темы, в частности, о поездке говорящего в Казань, последних новостях общих знакомых и комнатных растениях. Поскольку мы слышим только одну сторону разговора, звукозапись также представляет собой почти монологическую речь – истории, рассказанные говорящим, лишь иногда прерываются рядами частиц вроде *угу*, которые указывают на то, что в эти моменты она слушает своего собеседника. Однако в аудио есть небольшой отрывок, когда говорящий куда-то перемещает микрофон (возможно, просто трогает его, отвлекшись на разговор); в этой части ошибки распознавания становятся грубее. Большинство топонимов расшифровано верно, за исключением Свияжска – он превратился в просто *Яшск*. Упоминающаяся в разговоре *Зиночка* в некоторых местах в этой расшифровке тоже стала *Диночкой*, но в большинстве случаев была записана верно; *у Серёжки* превратилось в *а свёжки*. «Цветочная» лексика, в свою очередь, модели поддалась не вполне: *цвёл* во многих случаях заменено на *свёл*, *блёклые* (о цвете орхидей) – на *флёклые*, хотя остальные оттеночные прилагательные (*желтоватая*, *розоватая*) переданы верно. Ошибки в этой расшифровке по большей части малозаметные, например, *забрала время* вместо *забрал на время*, *отправдание* вместо *оправдание* и в падежных окончаниях (например, *декоративной* вместо *декоративная*).

Теперь перейдем к речевым эпизодам, которые были распознанные моделью хуже.

Для эпизода ordS126-14 WER составил 28%. Этот речевой эпизод представляет собой телефонный разговор, посвященный организации похорон. К возможным причинам ухудшения показателя WER можно отнести большое количество имен собственных при перечислении ожидаемых гостей.

Для эпизода ordS127-09 WER составил 41%. В этой расшифровке отсутствуют некоторые фразы, видимо, из-за колеблющегося уровня громкости голосов говорящих. В целом речевой эпизод представляет собой полилог с участием родите-

лей и детей, поэтому можно предположить, что сложности с его распознаванием связаны еще и с частой сменой говорящего.

В расшифровке эпизода ordS140-09 Whisper часто заменяет слова на похожие, например, *сытая пришла* на *сюда пришла*, *чашки били* на *чашки убери* или *надо* на *на дом*. Имена собственные модель передает не на 100% точно, но достаточно близко к их настоящему написанию: фамилия *Чубарова* была записана как *Чубарата*, *на Тореза* как *на Тараза*, тогда как более простые фамилии (*Сидоров*) и топонимы (*в Сосновке*) были расшифрованы правильно.

Далее обратимся к эпизодам, которые при помощи Whisper были распознаны хуже всего. Один из них – это речевой эпизод ordS11-13. WER для него составил 90%. Этот эпизод представляет собой разговор нескольких людей на рабочие темы. Можно заметить, что модель передала только часть реплик – в большинстве случаев те, которые произносит говорящий, находящийся ближе других к микрофону. Более того, когда ведущая роль в разговоре переходит ко второму говорящему, модель начинает «акцентировать внимание» только на его фразах.

В Таблице представлены некоторые типы ошибок, которые были выявлены при анализе других эпизодов.

Таблица

Типы ошибок, допускаемых Whisper при распознавании речи

Тип ошибки	Комментарии и примеры
«Орфографические» ошибки, ошибки в окончаниях	Небольшие ошибки есть практически в каждом документе с расшифровкой: *вести* вместо *ввести*, *ирфаке* вместо *юрфаке* (ordS27_27) и т. д.
Ошибки в именах собственных	Неточности самого разного характера, например: *Мороз* вместо *Морозов* (ordS16_17), *Вася* вместо *Васса* (ordS22_01) и т. д.
Ошибки в словах с уменьшительно-ласкательными суффиксами	Например: *Лосечка*, *лисичка* вместо *пусечка*, *кисичка* (ordS22_01).
Случайное повторение	Whisper имеет тенденцию «зацикливаться» и повторять фразу или словосочетание, которое говорящий не повторял (см. [Holtzman 2020]). Например, в конце эпизода ordS01-03 видим: *Да, правильно. Да, правильно. Да, правильно.*; тогда как на самом деле говорящий произнес эту фразу только один раз, что подтверждено и экспертной разметкой.
Пропуск слов	Более редкий вариант, в основном связанный с низкой громкостью.
Добавление слов	Более частый вариант: Whisper отделяет *-то* и некоторые приставки (особенно те, которые по форме совпадают с предлогами).
Замена слов на похожие	Замена не всегда происходит на полностью совпадающие по звучанию слова-омофоны, но на достаточно близкие: например, *проживает* вместо *прожевать* (ordS33_15).

Выводы

Таким образом, апробация модели Whisper, работающей за счет обучения на расшифровках на разных языках, на материале звукозаписей корпуса ОРД показала, что к ее достоинствам можно отнести качественную работу с частицами, топонимами и «проглоченными» слогами, однако она не всегда точно передает падежные и глагольные окончания. Кроме того, можно сделать предварительный вывод, что основными факторами, влияющими на качество автоматической расшифровки полевых записях с использованием модели Whisper, являются количество и смена говорящих, наличие фоновых шумов и громкость голоса говорящего, а также использование им специфической лексики.

Благодарности

Публикация подготовлена в результате проведения исследования по проекту «Текст как Big Data: моделирование конвергентных процессов в языке и речи цифровыми методами», выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2023 г.

Источники

Документация JiWER. [Электронный ресурс] URL: https://jitsi.github.io/jiwer/ (дата обращения: 11.10.2023).

Документация Whisper. [Электронный ресурс] URL: https://openai.com/research/whisper (дата обращения: 11.10.2023).

Список литературы

Богданова-Бегларян Н.В. и др. Корпус русского языка повседневного общения. Один речевой день (ОРД): текущее состояние и перспективы / Н.В. Богданова-Бегларян, О.В. Блинова, Г.Я. Мартыненко, Т.Ю. Шерстинова // Труды Ин-та русского языка им. В.В. Виноградова. 2019. № 3(21). С. 100–110. [Электронный ресурс]. URL: https://ruslang.ru/doc/trudy/vol21/5-bogdanova-beglaryan.pdf (дата обращения: 11.10.2023).

Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений симво-

лов // Доклады Академий наук СССР. 1965. Т. 163, № 4. С. 845–848.

Ali A., Renals S. Word Error Rate Estimation for Speech Recognition: e-WER // Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: Association for Computational Linguistics, 2018. Vol. 2: Short Papers. Pp. 20–24. [Электронный ресурс]. URL: https://aclanthology.org/P18-2004.pdf (дата обращения: 11.10.2023).

Asinovsky A. et al. The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation / A. Asinovsky, N. Bogdanova, M. Rusakova, A. Ryko, S. Stepanova, T. Sherstinova // Lecture Notes in Artificial Intelligence (LNAI). 2009. Vol. 5729. P. 250–257.

Gris L.R.S. et al. Evaluating OpenAI's Whisper ASR for Punctuation Prediction and Topic Modeling of Life Histories of the Museum of the Person / L.R.S. Gris, R. Marcacini, A. Candido Jr, E. Casanova, A. Soares, S.M. Aluísio. 2023. [Электронный ресурс]. URL: https://arxiv.org/abs/2305.14580 (дата обращения: 11.10.2023).

Holtzman A. et al. The Curious Case of Neural Text Degeneration / A. Holtzman, J. Buys, Li Du, M. Forbes, Yejin Choi // Proceedings of the 2020 International Conference on Learning Representations. 2020. P. 2540.

Lin H.-Y. Standing on the Shoulders of AI Giants. 2023. [Электронный ресурс] URL: https://www.computer.org/csdl/api/v1/periodical/mags/co/2023/01/10008980/1JIoHOJrayA/download-article/pdf (дата обращения: 11.10.2023).

McCowan I. et al. On the Use of Information Retrieval Measures for Speech Recognition Evaluation / I. McCowan, D.C. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, H. Bourlard // IDIAP Research Report. 2005. [Электронный ресурс]. URL: https://www.researchgate.net/publication/37433359_On_the_Use_of_Information_Retrieval_Measures_for_Speech_Recognition_Evaluation (дата обращения: 11.10.2023).

Radford A. et al. Robust Speech Recognition via Large-Scale Weak Supervision / A. Radford, J.W. Kim, Tao Xu, G. Brockman, C. McLeavey, I. Sutskever. 2022. [Электронный ресурс]. URL: https://cdn.openai.com/papers/whisper.pdf (дата обращения: 11.10.2023).

Sherstinova T. Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus // *SPECOM 2015.* Lecture Notes in Artificial Intelligence. 2015. Vol. 9319. Pp. 268–276.

THE WRITER ROBIN DRANATHTAGORE: APPROBATION OF THE WHISPER MODEL IN RUSSIAN SPEECH

Evgenia O. Kolpashchikova
Research Intern, Linguistic Convergence Laboratory
National Research University Higher School of Economics – St Petersburg

Whisper is an acoustic model released by OpenAI about a year ago. Whisper was trained on 680,000 hours of multilingual and multitasking speech, which should improve the model's performance in recognizing accents and make it less sensitive to background noise. The study is devoted to testing the capabilities of Whisper on field audio recordings from the "One Speaker's Day" corpus and assessing the accuracy of the resulting transcripts (compared to manual ones). The study also describes and groups the errors made by the model and identifies other features of its work: for example, Whisper quite accurately restores syllables "swallowed" by the speaker, but does not always correctly capture grammatical endings.

Keywords: automatic speech recognition; Russian language; everyday speech communication; automatic and expert transcription; Whisper.