УДК 81’33

КОЛЛАБОРАЦИЯ С ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ: ПОЛУАВТОМАТИЧЕСКАЯ РАСШИФРОВКА ЗАПИСЕЙ РЕЧИ УЧИТЕЛЕЙ [1]

Иван Дмитриевич Мамаев
ассистент кафедры математической лингвистики
Санкт-Петербургский государственный университет,
199034, Санкт-Петербург, Университетская наб. 7/9. i.mamaev@spbu.ru
старший преподаватель кафедры Р7 «Теоретической и прикладной лингвистики»
Балтийский государственный технический университет «Военмех» им. Д.Ф. Устинова
190005, Санкт-Петербург, 1-я Красноармейская 1. mamaev_id@voenmeh.ru

Елена Игоревна Риехакайнен
к. филол. н., доцент кафедры общего языкознания им. Л. А. Вербицкой
Санкт-Петербургский государственный университет
199034, Санкт-Петербург, Университетская наб. 7/9. e.riehakajnen@spbu.ru

Лада Леонидовна Шумакова
стажер-исследователь филологического факультета
Санкт-Петербургский государственный университет
199034, Санкт-Петербург, Университетская наб. 7/9. skorobagatko.lada@gmail.com

В статье представлен алгоритм расшифровки записей для корпуса устной речи учителей. Орфографическая расшифровка и фонетическая транскрипция осуществляются в полуавтоматическом режиме. Используются автоматические системы распознавания речи, результат работы которых затем корректируется лингвистами – разработчиками корпуса. Для орфографической расшифровки применяется программа Whisper, для фонетической транскрипции – специально разработанный для нашего корпуса алгоритм на базе инструмента для распознавания речи с открытым исходным кодом Kaldi. В статье рассматриваются типичные ошибки, которые возникают при автоматической расшифровке.

Ключевые слова: русская устная речь; автоматическое распознавание речи; транскрипция; Whisper; Kaldi.

Введение

Несмотря на то что в последние годы появляется все больше корпусов устной речи на материале различных языков, их количество и объем по-прежнему существенно уступают корпусам письменной речи. Это связано с несколькими причинами, но одной из основных, вне всякого сомнения, является то, что устный речевой сигнал нужно предварительно превратить в текст, т. е. расшифровать. По-видимому, в любом корпусе устной речи представлена орфографическая расшифровка. Насколько нам известно, во всех существующих общедоступных корпусах русской устной речи до недавнего времени она выполнялась экспертами-лингвистами вручную, что существенно замедляло процесс создания корпуса (о недавних экспериментах по применению автоматических систем при расшифровке материала для корпусов русской устной речи см., например, [Sherstinova et al. 2024]). В настоящее время мы разрабатываем корпус речи школьных учителей, материалы которого будут в дальнейшем использоваться в том числе для машинного обучения. Мы заинтересованы в получении большого массива качественно расшифрованных текстов и поэтому решили разработать систему полуавтоматической расшифровки, которая предполагает первичную обработку звукового сигнала с помощью автоматической системы распознавания речи и последующую проверку получившейся расшифровки экспертом. На основе проведенного нами ранее анализа того, как справляются с задачей распознавания естественной русской речи три автоматические системы, было решено остановиться на использовании системы Whisper. В следующем разделе статьи будет подробно описан алгоритм этого этапа работы над корпусом.

Помимо орфографической расшифровки, планируется сопроводить корпус акустико-фонетической транскрипцией, подобной той, которая применяется в Корпусе русской устной речи (russpeech.spbu.ru). Этот корпус на данный момент уступает по объему большинству других общедоступных корпусов русской устной речи, поскольку акустико-фонетическая транскрипция занимает намного больше времени, чем орфографическая расшифровка. В третьем разделе статьи будет представлен автоматический алгоритм, который, на наш взгляд, позволит ускорить и упростить и этот этап создания корпуса устной речи.

Полуавтоматическая орфографическая расшифровка записей

Whisper – это крупномасштабная автоматическая система распознавания речи, анонсированная компанией OpenAI в сентябре 2022 г. [Radford et al. 2023]. Она представляет собой модель, обученную на многомиллионном корпусе аудиозаписей на различных языках. Инструмент Whisper был создан с целью повышения точности распознавания речи в условиях реального мира: при шуме, наличии акцентов и низком качестве записи, что часто затрудняет процесс распознавания речи в традиционных системах. Whisper базируется на архитектуре трансформеров (Transformers), которая используется для обработки последовательностей данных. Она обучена на 680 тыс. часов данных, что обеспечивает высокую точность и адаптивность к различным условиям записи. Whisper использует несколько этапов преобразования и обработки аудиосигнала: преобразование аудио в мел-спектрограммы, введение последовательности в архитектуру трансформеров, в которой выполняется интерпретация речевого сигнала, декодирование данных в текст. Каждый из этих этапов оптимизирован для обеспечения устойчивости к шуму и искажениям. Сам инструмент уже апробирован на разножанровых корпусах русского языка [Колпащикова 2023; Amorese et al. 2023; Sherstinova et al. 2024].

На сегодняшний день Whisper предлагает несколько моделей разного размера и производительности. В настоящем исследовании для расшифровки текстов корпуса используется large-модель (1550M параметров), которая подходит для сложных задач распознавания спонтанной и подготовленной речи, а также для задач с высокой степенью многоязычности. Методика работы заключается в интеграции инструмента в рабочую среду Google Colab с получением расшифровки в txt-формате, которая в дальнейшем перепроверяется лингвистами-экспертами. Основные недостатки, которые были выявлены при тестировании этой системы на материале записей из Корпуса русской устной речи, описаны в [Мамаев, Риехакайнен 2023]. С учетом этих данных была составлена инструкция для экспертов, которые должны проверять автоматические расшифровки до их включения в корпус. Алгоритм полуавтоматической орфографической расшифровки был апробирован на материале 55 уроков школьных учителей – носителей русского языка (продолжительность каждого урока – около 40 мин.). В ходе предварительного анализа результатов апробации к наиболее частотным и критичным для наших целей были отнесены следующие особенности автоматических расшифровок:

- расшифровки приближены к письменной речи, т. е. система игнорирует повторы, самоисправления и другие речевые сбои, которые могут оказаться значимыми для создания речевого портрета учителя;
- часто возникают проблемы с расшифровкой имен собственных;
- не разделяется речь учителя и учеников;
- встречаются т. н. «зацикливающиеся» расшифровки – по какой-то причине автоматическая система не распознает часть звукового сигнала, а выдает в качестве результата многократно повторяющийся фрагмент текста.

До выяснения истинных причин возникновения последней проблемы единственным решением является полностью экспертная расшифровка таких фрагментов. Остальные сложности учитываются в инструкции, по которой работают эксперты, а именно – указывается не только необходимость исправления всех встретившихся в тексте ошибок и единообразного оформления орфографической записи (например, отражение «ё» в тексте, запись числительных словами), но и вводятся следующие пункты инструкции:

- делим расшифровку на речь учителя (Учитель:) и учеников (Ученик:);
- весь фрагмент до смены говорящего объединяем в один абзац;
- добавляем все то, что пропущено, включая повторы слов;
- используем многоточие без пробела после «оборванного слова» для самоисправлений и обрывов (*какими молф… морфологическими*);
- проверяем написание всех имен собственных, при необходимости исправляем;
- стараемся точно воспроизводить, что именно было произнесено в тех случаях, когда возможна вариативность типа *чтобы/чтоб*.

Таким образом за сравнительно короткое время мы получаем расшифровку, приближенную по качеству к расшифровке, которую выполняет эксперт, работающий без обращения к автоматическим системам распознавания речи.

Тестирование алгоритма автоматической акустико-фонетической транскрипции устной речи

Для создания алгоритма автоматического транскрибирования записей устной речи мы выбрали программу Kaldi (https://kaldi-asr.org). Эта программа имеет открытый исходный код, и в ней предусмотрена возможность обучать модели на собственном материале. При этом есть два уровня моделирования: акустический (извлечение акустических характеристик речи) и лингвистический (учитывает собственно языковые особенности и контекст). Эта программа была протестирована на материале нескольких языков и показала хорошие результаты (см., например: [Hui Bu et al. 2017: electr. resource; Kew et al. 2020; Linke et al. 2023: electr. resource]). Поскольку нашей задачей было получение транскрипционной записи, в качестве единиц словаря для тестирования моделей был использован словарь аллофонов, в который вошли все варианты фонем, различаемые при транскрибировании Корпуса русской устной речи: от набора фонем русского языка этот список отличается в первую очередь тем, что отдельными элементами считаются гласные звуки в зависимости от твердости/мягкости предшествующего и последующего согласных, а также отдельно отмечаются долгие звуки (подробнее см.: https://russpeech.spbu.ru/transkrip.htm). Следовательно, в качестве основной метрики для тестирования разработанного алгоритма использовался параметр PER (Phoneme Error Rate), который представляет собой выраженное в процентах отношение суммы количества замен, вставок и удалений аллофонов к общему числу аллофонов. Чем ниже значение этой метрики, тем лучше предсказание модели.

Нами были протестированы модели как с монофонным, так и с трифонным обучением, а также при варьировании объема контекста (N-грамм). В качестве материала использовались записи из Корпуса русской устной речи, потому что для них уже есть выполненная экспертами транскрипция. Это были 3 083 межпаузальных интервала из трех аудиофайлов разной тематики (общая продолжительность – 92 мин. 53 сек.). Все записи представляли собой спонтанные монологи (в том числе как часть диалога). Этот материал был разделен на обучающую и тестовую выборки в соотношении 8 к 2: 2 466 межпаузальных интервалов вошли в обучающую выборку и 617 – в тестовую. Наиболее успешным оказалось трифонное моделирование: при длине N-грамм 4 и более показатель PER равняется 36–37% (подробно результаты тестирования алгоритма описаны в [Скоробагатько 2024]).

Поскольку примерно треть аллофонов все-таки была распознана неверно даже в лучших модификациях модели, мы постарались определить, с чем связаны ошибки системы. Были выдвинуты следующие предположения: 1) модель лучше распознает те звуки, которые наиболее часто встречаются в рамках корпуса, потому что они имеют большую репрезентативность; 2) аллофоны гласных распознаются хуже, чем алллофоны согласных, поскольку в процессе расшифровки даже у экспертов возникали трудности при решении вопроса о том, как стоит классифицировать тот или иной гласный звук, а инструментальный анализ формантных характеристик не всегда позволял разрешить сомнения [Nigmatulina et al. 2016: 178–179].

Было решено использовать модели монофонного и трифонного обучения длиной 6, так как эта длина N-грамм оптимальна с точки зрения вычислительной мощности и, кроме того, существенного улучшения результатов при увеличении этого параметра не происходит.

Были проанализированы все аллофоны, встретившиеся в тестовой выборке не менее 50 раз, для них рассчитаны значения метрики Accuracy, которая представляет собой отношение правильно предсказанных единиц к их общему количеству (чем ближе значение Accuracy к единице, тем лучше предсказывается аллофон). Гипотеза о том, что согласные аллофоны будут распознаваться успешнее, чем гласные, подтвердилась: среди десяти звуков с наибольшим значением Accuracy для каждой из моделей только три являются гласными. При этом частота встречаемости аллофонов гласных иногда выше, чем у более удачно распознаваемых согласных, например: в пределах тестовой выборки [a] встречается 731 раз, но точность его распознавания ниже, чем у согласных [t], [ts] и [n’] при распознавании моделью, основанной на монофонном анализе. Среди десяти наименее успешно распознанных аллофонов шесть и восемь являются гласными (для монофонного и трифонного обучения соответственно). Встречаемость аллофонов в тестовой выборке не коррелирует с успешностью их распознавания при монофонном обучении ($r = 0,3$; $p = 0,54$), при трифонном обучении есть умеренная значимая корреляция ($r = 0,41$; $p = 0,008$).

Кроме того, для каждой модели были построены матрицы ошибок, в которых было описано,

какие звуки, каким образом и в каком количестве распознавались. На основе этих данных были сделаны следующие наблюдения:

1) гласные путаются между собой значительно чаще, чем согласные, что коррелирует с представленными выше данными;
2) модели склонны распознавать долгие варианты аллофонов (которые маркируются в транскрипции знаком «#») как обычные звуки;
3) чаще всего модели заменяли аллофоны гласных на [ə] и [ɨ], что может быть связано с их большей представленностью в обучающей или тестовой выборке;
4) во многих случаях путались аллофоны, зависящие от мягкости/твердости предшествующего согласного и маркированные знаками «:» (после мягкого согласного или между мягкими согласными) и «I» (перед мягким согласным);
5) модели заменяют мягкие согласные на парные им твердые и наоборот;
6) происходит смешение акустически сходных аллофонов согласных (например, звук [ts] часто заменяется на [t] или [s]);
7) звонкие согласные иногда распознаются как парные им глухие (например, [z] как [s]), при этом шумные согласные не заменяются на сонорные.

Заключение

Проведенное исследование показало, что расшифровки записей устной речи можно и целесообразно проводить в полуавтоматическом режиме. Автоматические системы распознавания устной речи типа Whisper выдают вполне приемлемый результат, поэтому степень участия эксперта и исправления, которые он должен вносить, зависят от задач конкретного корпуса. Так, в нашем корпусе важно максимально точно фиксировать, что именно произнес учитель, поэтому нам важно, например, восстанавливать повторы, которые совершил учитель, но проигнорировала автоматическая система. Если целью расшифровки является передача только основной сути устного сообщения, то, возможно, такие исправления не потребуются.

Осуществление автоматической акустико-фонетической транскрипции ожидаемо оказалось более сложной задачей. Как и в случае с экспертной оценкой, гласные вызывают больше затруднений, чем согласные, поэтому при осуществлении полуавтоматической аннотации материалов нашего корпуса на этом уровне будем придерживаться того же принципа, к которому пришли ранее для экспертного транскрибирования [Риехакайнен и др. 2024: 276]: указываем все согласные и ударные гласные, для безударных гласных фиксируем только их наличие. Необходимо также отметить, что на данный момент алгоритм полуавтоматического транскрибирования протестирован только на материале Корпуса русской устной речи. Ближайшей задачей является проверка этого алгоритма на записях речи учителей из разрабатываемого корпуса.

Примечание

[1] Работа выполнена при поддержке СПбГУ, шифр проекта 103923108, и в рамках договора между СПбГУ и ООО «СберОбразование» № 230712-107-ЮЛ. Выражаем благодарность нашим коллегам по проекту Ю.С. Виноградовой, Е.С. Затеваловой, М.А. Осадчей, В.О. Прокаевой, У.А. Судаковой, которые принимали активное участие в разработке инструкции для полуавтоматической орфографической расшифровки речи учителей.

Список литературы

Колпащикова Е.О. Писатель Робин Дранаттагор: апробация модели Whisper на русскоязычной звучащей речи // Социо-и психолингвистические исследования. 2023. Вып. 11. С. 23–27.

Мамаев И.Д., Риехакайнен Е.И. Автоматическая расшифровка записей устной речи: тестирование программы Whisper // Социо- и психолингвистические исследования. 2023. Вып. 11. С. 19–22.

Риехакайнен Е.И. и др. Методика аннотирования корпуса устной речи учителей / Е.И. Риехакайнен, В.С. Браташ, В.И. Зубов, П.А. Сергоманов // Вопросы образования. 2024. № 2. С. 251–285.

Скоробагатько Л.Л. Автоматическое транскрибирование русской устной речи при помощи инструмента Kaldi // Фонетический лицей. СПб.: Скифия-принт, 2024. Вып. 9. С. 69–74.

Amorese T. et al. Automatic speech recognition (ASR) with Whisper: Testing performances in different languages / T. Amorese, C. Greco, M. Cuciniello, R. Milo, O. Sheveleva, G. Glackin // S3C'23: Sustainable, Secure, and Smart Collaboration Workshop Proceedings. 2023. Pp. 1–8.

Hui Bu et al. AISHELL-1: An open-source Mandarin speech corpus and a speech recognition baseline / Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, Hao Zheng // 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA). 2017. [Electronic resource]. URL: https://arxiv.org/abs/1709.05522 (date of access: 01.11.2024).

Kew T. et al. UZH TILT: A Kaldi recipe for Swiss German Speech to Standard German text / T. Kew, I. Nigmatulina, L. Nagele, T. Samardzic //

Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects. Barcelona: ICCL, 2020. Pp. 15–24.

Linke J. et al. Using Kaldi for automatic speech recognition of conversational Austrian German / J. Linke, S. Wepner, G. Kubin, B. Schupple. 2023. [Electronic resource]. URL: https://arxiv.org/abs/23-01.06475 (date of access: 01.11.2024).

Nigmatulina Ju. et al. How to study spoken word recognition: Evidence from Russian / Ju. Nigmatulina, O. Rajeva, E. Riechakajnen, N. Slepokurova, A. Vencov // Slavic Languages in Psycholinguistics: Chances and Challenges for Empirical and Experimental Research / T. Anstatt, A. Gattnar, C. Clasmeier (eds.). Tuebingen: Narr Francke Attempto Verlag, 2016. Pp. 175–190.

Radford A. et al. Robust speech recognition via large-scale weak supervision / A. Radford, Jong Wook Kim, Tao Xu, G. Brockman, C. McLeavey, I. Sutskever // Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR, 2023. Pp. 28492–28518.

Sherstinova T. et al. Bridging gaps in Russian language processing: AI and everyday conversations / T. Sherstinova, N. Mikhaylovskiy, E. Kolpashchikova, V. Kruglikova // 35th Conference of Open Innovations Association (FRUCT). IEEE, 2024. Pp. 665–674.

COLLABORATION WITH AI: SEMI-AUTOMATIC ANNOTATION OF TEACHERS' SPEECH

Ivan D. Mamaev
Assistant Lecturer, Department of Mathematical Linguistics
Saint Petersburg State University
Senior Lecturer, Department of Theoretical and Applied Linguistics
Baltic State Technical University "Voenmeh" named after D.F. Ustinov

Elena I. Riekhakaynen
Associate Professor, Department of General Linguistics
Saint Petersburg State University

Lada L. Shumakova
Research Assistant, Philological Faculty
Saint Petersburg State University

The paper describes an algorithm for transcribing recordings for a corpus of teachers' speech. Both orthographic annotation and phonetic transcription are carried out in semi-automatic mode. Automatic speech recognition systems are used, the result of which is then corrected by linguists – the developers of the corpus. For orthographic annotation, the Whisper ASR is used. For phonetic transcription, we developed an algorithm based on the open-source toolkit Kaldi for speech recognition. The article discusses typical errors that occur during automatic annotation and transcription.

Keywords: Russian speech; automatic speech recognition; transcription; Whisper ASR; Kaldi ASR.