УДК 81’33

АВТОМАТИЧЕСКАЯ РАСШИФРОВКА ЗАПИСЕЙ УСТНОЙ РЕЧИ: ТЕСТИРОВАНИЕ ПРОГРАММЫ WHISPER[1]

Иван Дмитриевич Мамаев
ассистент кафедры математической лингвистики
Санкт-Петербургский государственный университет,
199034, Санкт-Петербург, Университетская наб., 7/9, i.mamaev@spbu.ru
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

В статье представлено сопоставление результатов расшифровки записей из Корпуса русской устной речи (russpeech.spbu.ru), выполненной экспертами и автоматической системой распознавания речи Whisper. Для формальной оценки качества автоматических расшифровок использовались русскоязычные языковые модели типа Transformers. Средний коэффициент косинусной близости между эталонными ручными расшифровками и созданными автоматически текстами равен 0,8. Большая часть ошибок, совершенных автоматической системой, не является критичными. Полученные результаты указывают на относительно высокую степень сохранности семантики исходной записи и на потенциальную возможность использования программы Whisper при расшифровке записей для корпуса речи школьных учителей, который мы разрабатываем в настоящее время.

Ключевые слова: автоматическое распознавание речи; русская устная речь; корпусное исследование; Whisper.

1. Введение

В настоящее время технологии автоматического распознавания речи хорошо справляются с прикладной задачей расшифровки речевого сигнала и все шире используются на практике. Вместе с тем при создании корпусов устной речи лингвисты по-прежнему предпочитают осуществлять расшифровку вручную: именно так, например, созданы корпусы, представленные на сайте «Рассказы о сновидениях и другие корпуса звучащей речи» (http://spokencorpora.ru/), корпус «Один речевой день» (https://ord.spbu.ru), Корпус русской устной речи (http://russpeech.spbu.ru). Экспертная расшифровка является время- и ресурсозатратной, поэтому корпусы устной речи, как правило, существенно уступают корпусам письменной речи по объему.

В нашем проекте разрабатывается корпус речи учителей, чтобы на его основе выявить лингвистические характеристики, отличающие результативные учительские практики от менее результативных. Очевидно, что для решения этой задачи необходим большой объем расшифрованного материала, который впоследствии может быть проанализирован в корпусных исследованиях и привлечен в качестве источника материала для экспериментов по восприятию речи. Мы предположили, что процесс обработки материала может ускориться, если использовать для первичной орфографической расшифровки автоматическую систему распознавания устной речи и затем вручную корректировать результаты ее работы. В статье описывается проведенное нами тестирование одной из таких систем.

2. Методика и материал исследования

Мы использовали систему автоматического распознавания речи с открытым кодом Whisper (разработка компании Open AI; см. подробнее https://openai.com/research/whisper), обученную на сверхбольшом наборе многоязычных данных, которые включают и русскоязычный сегмент. По данным ряда исследований [Dewan et al. 2023: электр. ресурс; Junczyk 2023; Radford et al. 2023: электр. ресурс и др.], эта программа показывает

стабильно хорошие результаты по распознаванию речи на разных языках. То, что программа успешно справляется с распознаванием речи на разных языках (а не только на русском), может в перспективе оказаться важным для нашего исследования, поскольку в корпус речи учителей могут быть включены уроки иностранного языка в школе, во время которых и учителя, и ученики, как правило, говорят и на родном языке, и на иностранном. У программы Whisper существует несколько вариантов; в нашем исследовании мы использовали модель Whisper large-v2, которая обычно показывает более высокие результаты, чем все остальные модели в рамках этой программы (см. сравнение разных моделей программы, например, в [Radford et al. 2023: электр. ресурс и др.]).

Методика исследования заключалась в сравнении орфографических расшифровок, выполненных экспертами-лингвистами и программой, и в лингвистическом анализе выявленных расхождений. В качестве материала были выбраны три аудиозаписи из Корпуса русской устной речи общим объемом более 14 тыс. словоформ (общая длительность записей – 92 мин. 52 сек.). В одной из записей было представлено радиоинтервью, в котором принимали участие журналист и гость студии, две остальные записи – телевизионные ток-шоу. Мы выбрали именно эти записи, потому что у них хорошее качество и это образцы естественной речи. Как и для всех текстов, которые входят в Корпус русской устной речи, для этих записей есть не только орфографическая расшифровка, но акустико-фонетическая транскрипция. Следовательно, в тех случаях, когда обнаруживаются несоответствия экспертной орфографической расшифровки и результатов автоматического распознавания звукового сигнала, можно по транскрипции проверить, не связано ли это с несовершенством звукового сигнала (с наличием редукции, стяжений и т. п.).

3. Принципы анализа результатов

Первый этап оценки качества расшифровок включал формальную оценку семантической близости автоматически расшифрованных текстов и эталонных размеченных текстов. Мы исходим из предположения, что значения косинусного сходства двух векторных представлений текстов, близких к единице, могут стать основанием для дальнейшей ручной экспертной оценки. В связи с активным применением нейросетевых архитектур Transformers в лингвистике [Blinova, Tarasov 2022; Fenogenova 2021; Ivanov 2022] для сравнения текстов был использован фреймворк SentenceTransformer и модель symanto/sn-xlm-roberta-base-snli-mnli-anli-xnli, обученную для задач семантических измерений на части русскоязычных данных. Средний коэффициент косинусной близости между эталонными ручными расшифровками и автоматическими текстами оказался равен 0,8, что указывает на относительно высокую степень качества сохранения семантики исходной записи и возможность проведения дальнейшей экспертной оценки.

4. Результаты

Прежде всего нужно отметить, что в некоторых случаях в расшифровках, выполненных автоматической системой, содержится больше информации, чем в расшифровках экспертов. В первую очередь это относится к случаям, когда два или более дикторов говорят одновременно. При ручной расшифровке эксперты обычно отмечали такие фрагменты как «ансмбл» или «нрзб» («неразборчиво») и не расшифровывали, потому что предполагалось, что для этих случаев невозможно будет сделать акустико-фонетическую транскрипцию речи каждого из дикторов. Автоматическая же система выдает некоторый вариант распознавания речи одновременно говорящих людей. С точки зрения создания корпуса речи учителей, расшифровка одновременного говорения нескольких дикторов может быть полезной, поскольку на уроке нередко встречаются ситуации, когда учитель говорит на фоне речи кого-то из учеников. Вместе с тем в большинстве случаев результат расшифровки этих фрагментов программой оказался не вполне корректным (что связано, в том числе с тем, что программа не осуществляет распределение речи по дикторам).

Все ошибки, совершенные программой Whisper, были разделены на два типа: некритичные и критичные. Это деление опирается на представление о том, что не все ошибки в одинаковой мере влияют на успешность коммуникации, и сходно с принятым в методике преподавания русского языка как иностранного разделением ошибок на коммуникативно значимые (приводящие к коммуникативному сбою, т. е. «изменению/искажению намерений говорящего» [Косарева, Лазарева, Лужковская 2005: 4]) и коммуникативно незначимые (не вызывающие подобных изменений/искажений).

В расшифровках, выполненных программой, некритичных ошибок оказалось больше, чем критичных (180 и 116 соответственно). Среди ошибок первого типа можно выделить следующие группы (далее по тексту будут приведены бинарные примеры вида «ручная_расшифровка – автоматическая_расшифровка», знаки препинания в расшифровках не ставятся):

– вариативность частиц и союзов (*же – ж*; *чтоб – чтобы*);

– изменение краткой формы прилагательного на полную и наоборот (*не развиты – не развитые*; *все будут воспитанные – все будут воспитаны*);
– взаимозамены форм повелительного и изъявительного наклонений (*вы депутат вот вы и **придумайте** – вы депутат вот вы и **придумаете***; *...как угодно это **называйте** чтоб это были знания… – ...как угодно это **называете** чтоб это были знания…*);
– изменение числа или рода существительных, не приводящее к критичному нарушению общего смысла высказывания (*школа обстругивает детей потому что у них нет **ресурса** у этих учителей – школа обстругивает детей потому что у них нет **ресурсов** у этих учителей*; *учитель который сидит после **уроков** к которому можно прийти обратиться за консультацией – учитель который сидит после **урока** к которому можно прийти обратиться за консультацией*);
– нарушение согласования, которое восстанавливается из контекста (*мы столкнулись ещё с институтом преподавателей, **которые** вот сохраняли ту прежнюю традицию… – мы столкнулись ещё с институтом, преподавателей **которого** сохраняли эту прежнюю традицию…*);
– опущение заполнителей пауз (*а*, э);
– опущение слов, которые диктор повторяет в потоке речи (*это дело частной **частной** семьи – это дело частной семьи*; *а вот и **и** так далее – а вот и так далее*).

К типичным критичным ошибкам мы отнесли следующие:
– ошибки в распознавании имен собственных (*Савченко – Сапченко*; *Лукоморье – Рукоморье* и др.);
– неверное распознавание имен числительных (*два пятьдесят – двести пятьдесят*; *где-то четыреста долларов, {в} Москве там **восемьсот – тысячу**… – где-то 400 $ в Москве, там **800000**…*; *восемьсот двенадцатый год – восемнадцатый год*);
– неверное распознавание глаголов (*и того места где **находится** уважаемый депутат – того места где **был** уважаемый депутат*);
– опущение ряда придаточных предложений или целых предложений (*а вот собственно воспитание как функция школы а это в чём **это игра зарница это комсомол это классные собрания это что такое** это работа с отстающими после уроков»* – жирным выделен фрагмент, не распознанный программой).

5. Обсуждение результатов

В целом, анализ показывает, что критичных ошибок, во-первых, меньше, чем некритичных, во-вторых, они менее разнообразны (и по сути – за несколькими единичными исключениями – могут быть сведены к перечисленным выше типам). Причиной ошибок в распознавании числительных и глаголов можно считать фонетическую редукцию соответствующих словоформ. Так, в приведенных выше примерах глагол *находится* был произнесен диктором как [naxoetsʲ], а числительное *восемьсот двенадцатый* как [vosʲmsodʲunatstɨ]. Ошибки же в именах собственных, которые являются самыми частотными в проанализированном материале, объясняются, по всей видимости, отсутствием соответствующих единиц в словаре, с которым программа сравнивает входной речевой сигнал.

В речи учителей на уроке имена собственные и числительные встречаются достаточно часто, что может негативно сказаться на успешности автоматической расшифровки записей школьных уроков. Вместе с тем можно ожидать, что распространенные имена собственные будут распознаваться хорошо (в проанализированных записях такие имена собственные, как *Гарри Поттер*, *Евгений Герасимов*, *Андрей*, *Микеланджело* и пр., распознавались верно). Успешность же распознавания числительных будет зависеть от того, насколько четко говорит конкретный учитель.

При сопоставлении расшифровок мы относили пропуски повторяющихся слов к некритичным ошибкам, поскольку они не меняют смысл высказывания. Однако с точки зрения анализа педагогического дискурса подобные расхождения расшифровки с исходной записью могут оказаться критичными: повтор слов может быть приемом, который учитель осознанно или неосознанно использует, чтобы ученики лучше усвоили материал или для каких-то иных целей (например, в случае хезитации). Поэтому для выявления особенностей лингвистической организации речи учителя на уроке важно как можно более полно передать в расшифровке все, что было произнесено.

6. Заключение

Анализ результатов автоматического распознавания программой Whisper трех записей из Корпуса русской устной речи показал, что в целом программа успешно справилась с расшифровкой: бóльшая часть ошибок относится к некритичным, что позволяет использовать эту программу для первичной расшифровки записей при создании корпуса речи учителей. Проведенный анализ ошибок позволил определить, на какие именно фрагменты речевого сигнала нужно будет в первую очередь обращать внимание эксперту, который будет осуществлять проверку и доработку расшифровки, полученной автоматически.

Мы полагаем, что для улучшения работы программы ее необходимо тренировать на больших объемах спонтанной речи, что должно позволить избежать хотя бы части ошибок, связанных с неверной интерпретацией редуцированных словоформ.

Наша дальнейшая работа по переходу на автоматическую орфографическую расшифровку записей для корпуса будет включать в себя, с одной стороны, тестирование того, как программа Whisper будет справляться с расшифровками именно школьных уроков, а с другой – поиск других программных решений этой задачи. В частности, среди программ с открытым кодом, которые, по имеющимся в литературе данным [Долженко, Школина 2022], хорошо справляются с речью на русском языке, упоминается программа Vosk (https://alphacephei.com/vosk), которую мы планируем протестировать на имеющемся у нас материале в ближайшее время.

Примечание

[1] Исследование выполняется в рамках проекта СПбГУ (ID 101747352) и договора с ООО «СберОбразование» № 230712-107-ЮЛ.

Список литературы

Долженко А.И., Школина А.В. Обзор существующих систем распознавания речи с открытым исходным кодом // Проблемы проектирования, применения и безопасности информационных систем в условиях цифровой экономики: материалы XXII Международной научно-практической конференции. Ростов/Д: Издательско-полиграфический комплекс Ростовского государственного экономического университета (РИНХ), 2022. С. 341–345.

Косарева Е.В., Лазарева О.А., Лужковская М.Ф. Лингводидактическое тестирование: Процедура и методика проведения тестирования в рамках российской государственной системы тестирования. СПб.: Филологический факультет СПбГУ, 2005. 106 с.

Blinova O., Tarasov N. A Hybrid Model of Complexity Estimation: Evidence from Russian Legal Texts // Frontiers in Artificial Intelligence. 2022. Vol. 5. Pp. 1–14.

Dewan A. et al. Developing Automatic Verbatim Transcripts for International Multilingual Meetings: An End-to-End Solution / A. Dewan, M. Ziemski, H. Meylan, L. Concina, B. Pouliquen // MT Summit. Macau, 2023. [Электронный ресурс]. URL: https://arxiv.org/ftp/arxiv/papers/2309/2309.15609.pdf (дата обращения: 25.10.2023).

Fenogenova A. Russian Paraphrasers: Paraphrase with Transformers // Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing. Kiyv: Association for Computational Linguistics, 2021. Pp. 11–19.

Ivanov V.V. Sentence-Level Complexity in Russian: An Evaluation of BERT and Graph Neural Networks // Frontiers in Artificial Intelligence. 2022. Vol. 5. Pp. 1–12.

Junczyk M. BIGOS – Benchmark Intended Grouping of Open Speech Corpora for Polish Automatic Speech Recognition // Proceedings of the 18th Conference on Computer Science and Intelligence Systems. Annals of Computer Science and Information Systems. 2023. Vol. 35. Pp. 585–590.

Radford A. et al. Robust Speech Recognition via Large-Scale Weak Supervision / A. Radford, J.W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever // Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR, 2023. Pp. 28492–28518. [Электронный ресурс]. URL: https://proceedings.mlr.press/v202/radford23a.html (дата обращения: 25.10.2023).

AUTOMATIC SPEECH RECORDINGS RECOGNITION: TESTING WHISPER ASR

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

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

The article presents a comparison of the orthographic transcripts of recordings from the Corpus of Spoken Russian (russpeech.spbu.ru), performed by experts and the Whisper automatic speech recognition system. To formally assess the quality of automatic transcriptions, Russian-language models such as Transformers were used. The average cosine similarity coefficient between the manual and automatic transcripts is 0.8. Most of the errors made by the automatic system are not critical. The results obtained indicate a relatively high degree of preservation of the semantics of the original recording and the potential possibility of using the Whisper program when transcribing recordings for the corpus of speech of schoolteachers, which we are currently developing.

Key words: automatic speech recognition; Russian speech; corpus-based study; Whisper ASR.