

Автоматическое транскрибирование незнакомых слов для улучшения распознавания терминов предметной области

Волков Евгений Владимирович¹ (Evgenii V. Volkov),
Добров Борис Викторович² (Boris V. Dobrov)

¹ НИВЦ МГУ Имени М.В. Ломоносова, 119234, Россия, Москва, Ленинские Горы ул.,
д.1, стр.4

zhenia-465@mail.ru

² НИВЦ МГУ Имени М.В. Ломоносова, 119234, Россия, Москва, Ленинские Горы ул.,
д.1, стр.4,

dobrov_bv@mail.ru

Abstract. Одним из популярных методов улучшения качества распознавания незнакомых модели (OOV) слов является метод расширения словаря (vocabulary expansion). OOV слова особенно часто встречаются в предметных областях: это термины и терминоподобные словосочетания. В русской речи многие из них недавно заимствованы из английского языка и еще не имеют принятой записи на русском языке. Это затрудняет добавление таких слов в словарь модели для моделей, не являющихся мультязычными, что снижает качество их работы на аудио из предметных областей. В данной работе представлен метод улучшения качества распознавания речи, содержащей такие OOV термины, на основе алгоритма автоматического построения для произвольного английского слова т.н. русских транскрипций. Русская транскрипция – это сходно звучащее и записанное русскими буквами слово, которое могло бы заменить исходный термин при добавлении его в словарь модели, и таким образом сделать возможным его корректное распознавание. Результаты проведенных экспериментов позволяют говорить о возможности применения описанного алгоритма для улучшения качества работы ASR систем.

Keywords: ASR, OOV слова, транскрибирование, расширение словаря, предметная область.

1 Введение

С внедрением умных устройств и голосовых помощников задача автоматического распознавания речи (англ. ASR – Automatic Speech Recognition) становится все более актуальной. Разнообразие предлагаемых готовых решений способно удовлетворить запросы самой разной целевой аудитории, предлагая программные продукты для различных языков, задач, кошельков и различной степени доступности. Однако, практика показывает, что иногда результаты работы оказываются недостаточно точными при работе с не частотными словами, свойственными предметным областям – специальной терминологией, сленгом, неологизмами, жаргонизмами, именами собственными, эрративами и т.д. Поскольку ASR системы традиционно использовали словари для идентификации слов, такую лексику называют несловарными (OOV) словами (от англ. out of vocabulary – отсутствующие в словаре). Большинство современных моделей обычно обучается на больших, специально для этого созданных, корпусах данных. Такие корпуса обычно состоят из аудио общей направленности. Поскольку OOV слова редко (если вообще) встречаются в обычных текстах обуча-

ющей выборки, модель скорее всего не будет знакома с такими словами и не всегда будет способна их корректно распознавать.

Данная проблема является коммерчески актуальной для компаний, которые используют системы распознавания речи в своей деятельности. Поэтому для различных ситуаций были предложены различные методики работы с подобной лексикой [11], [16], [14]. В связи с гибкостью и низкой стоимостью [29] выделяются методы, основанные на расширении словаря (англ. vocabulary expansion). Они основаны на получении списка слов, которые модель должна «выучить». Сами слова обычно автоматически извлекаются из текстов, релевантных предметной области, в которой должна будет «разбираться» модель. С другой стороны, эти методы достаточно эффективны и позволяют «настраиваться» на предметную область довольно точно, что обеспечивает их популярность [10], [20].

Однако, нередко применение основанных на расширении словаря методов бывает ограничено особенностями внутренней реализации ASR моделей. В частности, одним из таких ограничений у многих не-мультиязычных моделей является возможность добавления в словарь только слов, записанных символами того есте-

ственного языка, с которым работает модель. С другой стороны, многие новые (и потому являющиеся OOV) слова приходят к нам из других языков, в современном русском языке чаще всего – из английского. Для них не существует принятой записи на языке, отличном от оригинала. Это ограничение является проблемным для тех языков, которые используют системы записи, отличные от латиницы, в частности для русского языка. В связи с этим, отсутствует возможность автоматизированного добавления слов, недавно пришедших из других языков (а также аббревиатур и имен собственных), в словари таких систем, что ограничивает качество их работы.

Поэтому перспективной видится разработка метода, который, используя сравнительно небольшие ресурсы, позволил бы мультязычным моделям работать с новыми англоязычными словами. Применение такого метода позволило бы улучшить качество распознавания путем переиспользования уже существующих методик пополнения словаря, что является ресурсоэффективным по сравнению с разработкой новой [29], [10] или дообучением на корпусе аудио старой модели [29].

В данной статье мы представляем алгоритм специального транскрибирования слов английского языка. Цель алгоритма – для англоязычного термина автоматически построить сходно звучащее русскоязычное слово, которое могло бы заменить исходный термин во внутреннем представлении модели. Для краткости мы будем называть такое слово русской транскрипцией.

Также описан механизм применения этого алгоритма для улучшения качества работы модели распознавания речи на примере модели VOSK 0.22 ru [24]. Выбор модели объясняется удобством применения в данной задаче: она не является мультязычной и, при этом, обладает достойным качеством.

Основной альтернативой предложенного подхода является использование мультязычных end-to-end (E2E) моделей, не имеющих промежуточного представления, и потому способных работать с иноязычными словами без необходимости адаптации написания этих слов к русскому языку. Однако, такие модели устроены значительно сложнее, а также не лишены других своих недостатков (см. главу 3.1).

2 Постановка задачи

2.1 Цель работы

Цель данной работы – разработка алгоритма транскрибирования (преобразования) англоязычных слов в т.н. русские транскрипции – слова, предназначенные для дальнейшего использования в процессе дообучения не являющихся мультязычными (моноязычных) моделей распознавания речи методом расширения словаря модели, а также метода применения разработанного алгоритма для повышения качества распознавания речи. При этом вышеупомянутые русские транскрипции должны обладать следующими свойствами:

- записываются буквами русского языка;
- имеют произношение, которое наиболее близко к произношению исходного англоязычного слова носителем русского языка;
- из-за разных вариантов произношения, для некоторых англоязычных слов, возможно, может существовать более одной соответствующей ему русской транскрипции;
- могут быть в автоматизированном режиме составлены для произвольного англоязычного слова, в том числе, возможно, отсут-

ствующего в словарях языка или не существующего на самом деле.

Используя разработанный алгоритм для построения русских транскрипций, оценить его эффективность путем проведения экспериментов над выбранной моделью.

В качестве модели предлагается использовать модель VOSK 0.22 ru. Эффективность предлагается оценивать как разность значений выбранной метрики качества при распознавании заданного набора аудио с использованием словаря, содержащего русские транскрипции, и без него.

Также требуется оценить влияние параметров алгоритма транскрибирования на изменение качества распознавания.

2.2 Актуальность

Во многих современных областях знаний, особенно высокотехнологичных, англоязычными неологизмами, не встречающихся в словарях русского языка, нередко является специальная лексика — термины, профессионализмы, имена собственные (имена, фамилии, названия компаний). В аудио, относящихся к различным

предметным областям знаний такие слова составляют саму семантику аудио. Многие из этих областей стремительно развиваются. Поэтому аудио, содержащие такие слова, могут содержать ценную информацию, не представленную ранее в печатных изданиях на русском языке, и получение транскриптов может стать источником такой информации в письменном виде. Некорректная же работа с такими словами может значительно повредить или исказить семантику аудио и сделать затруднительной обработку содержащейся в нем информации.

Поэтому можно сделать вывод, что рассматриваемая проблема распознавания незнакомых англоязычных слов может являться достаточно актуальной. Невозможность работы некоторых моделей с такой лексикой является серьезным фактором, значительно ограничивающим их применение, в том числе в области извлечения информации.

Возможность преодолеть эту проблему, используя довольно ограниченные ресурсы, также может считаться актуальной задачей, поскольку построение новых или дообучение на корпусах аудио старых моделей является очень ресурсоемкой задачей [29], [10].

Использование же мультязычных моделей не всегда возможно или оправдано из-за наличия проблем, вытекающих из их высокой сложности. К тому же, далеко не все представленные на международном рынке модели обладают достаточно высоким качеством модели для русского языка. И даже для таких моделей источником ошибок в билингвальных текстах может стать сильный акцент. Модель ожидает произношение слов по правилам языка, но большинство людей не является билингвами, и потому те же слова в их исполнении могут звучать совсем по-другому. Поэтому даже для мультязычных моделей предложенный метод может иметь потенциал применения в случае обилия спикеров с сильным акцентом или людей, произносящих английские слова не по правилам английского языка.

3 Обзор существующих решений

3.1 Мультязычные модели

Основной альтернативой подхода, заключающегося в разработке алгоритма транскрибирования, стоит считать такой подход, который позволил бы вообще не столкнуться с проблемой записи иноязычных слов. Действительно, существуют end-to-end (E2E) моде-

ли, многие из которых являются мультязычными. Эти модели не имеют промежуточного представления, и потому способны работать с иноязычными словами в исходной записи, без необходимости построения русской транскрипции для обработки или добавления в пользовательский словарь. И хотя в последнее время E2E модели уже практически вытеснили все остальные в повседневных приложениях [18], они имеют множество недостатков, ограничивающих эффективность их использования именно в задачах, связанных с OOV.

Качество (выраженное метрикой WER [25]) лучших E2E моделей значительно выше качества лучших не-E2E моделей, к которым относятся модели на основе скрытых марковских моделей (англ. HMM – hidden Markov model)). Однако, это достигается по большей части не за счет более богатого «лексикона», а за счет более качественной работы с повседневной лексикой (флексивной, омонимичной, нечетко произнесенной и т.д.), для работы с которой эти модели и предназначены. Таким образом, хотя лучшие E2E модели (в частности, Whisper [15]) на выходе способны давать грамматически связный русский текст с минимумом ошибок в по-

вседневной речи, проблемы с OOV лексикой для таких моделей никуда не исчезают.

Напротив, проблема OOV для них также является актуальной [26]. Применение E2E моделей для решения проблемы OOV сталкивается со значительными трудностями. Также многие популярные E2E модели разработаны в англоязычном мире, и именно английский язык является для них основным. Поэтому качество работы русской модели у многих крупных компаний (в частности, Google [9], Amazon [3], Microsoft [22]) является неудовлетворительным [2], [19], а многие модели, в том числе принадлежащие российским компаниям (в частности, у Silero [21] – одного из лидеров среди российских компаний [2]), могут быть закрыты для улучшения извне или не предназначены для использования индивидуальными исследователями, что значительно уменьшает количество доступных решений. Также дообучение таких моделей может являться технически более сложным процессом по сравнению с моделями, работающими со словарем.

Это, в частности, относится к Whisper. Хотя эта модель легко развертывается и обладает высоким качеством «из коробки», ее автоматизированная адаптация под предметную область достаточно

затруднительна. Whisper на данный момент является лидером с точки зрения качества распознавания речи, по крайней мере, на англоязычной речи [8]. Для русского языка найти комплексные и актуальные сравнения WER моделей, включающие Whisper, не удалось, однако проведенные авторами исследования позволяют говорить о высоком качестве работы Whisper для русского языка. Поэтому стоит коротко перечислить обнаруженные в процессе взаимодействия с этой моделью основные недостатки, осложняющие ее применения в задаче распознавания OOV:

- исправление некоторых ошибок, свойственных устной речи: грамматических (согласование родов, падежей), оговорок, повторов текста;
- оборванные на полуслове предложения (в частности, обрезанные предложения в конце аудио) автоматически достраиваются «фантазией» модели;
- сложно понять, какие слова реально для модели являются OOV, так как незнакомые слова модель пытается угадать и записать по звучанию, получая целую россыпь вариантов в одном и том же документе и иногда даже изобретая несуществующие слова:

— «HighLoad» -> «HiLot», «хайло», «high load»;

— «GitOps» -> «GitOps», «GeTops»;

- некоторые давно пришедшие и закрепившиеся в русском языке англоязычные слова и жаргонизмы могут вдруг быть расшифрованы как английские:

— «для пиара» -> «для Pearo»;

— «тимлид» -> «Team Lead»;

- часть слов записывается не словами, а числами или другими символами (например, символ процента «%»), теряя падежи и усложняя автоматизированные проверки качества.

-

- В общем, процесс работы и обучения сложных мультязычных моделей не всегда является прозрачным, а результат – предсказуемым, что делает их не слишком удобными в работе с OOV.
- С другой стороны, поскольку НММ модели работают на основе словаря, они могут быть адаптированы для использования в предметных областях малыми силами, путем расширения словаря модели словами из пользовательского словаря. Расширение словаря возможно в автоматизированном режиме, поскольку НММ модель, используя внутренние инструменты, может из-

влечь всю необходимую ей информацию об OOV слове из текста предметной области, содержащего это слово, без необходимости в аудио. Это означает отсутствие необходимости сбора корпусов релевантного аудио, обычно используемых для дообучения E2E моделей, что влечет за собой существенно меньшие трудозатраты для адаптации модели под предметную область. Поэтому использование продвинутых НММ моделей для извлечения OOV лексики для обучения больших E2E моделей может быть потенциально перспективным направлением.

- Таким образом, несмотря на значительное отставание в качестве, НММ модели все еще имеют ниши, в которых они могут являться достаточно актуальными. Поэтому составление методов, которые позволили бы повысить качество работы таких моделей, также остается актуальной задачей.

3.2 Существующие методы построения русских транскрипций

Задача построения русских слов, звучащих аналогично английским, является достаточно узкоспециализированной.

Это легко объясняется ее применимостью. Вероятность того, что носителю русского языка потребуется взаимодействие с английским языком несравнимо выше того, что носителю английского языка потребуется взаимодействие с русским. Поэтому, в отличие от обратной задачи английской транслитерации русских слов, которая, например, имеет применение даже в повседневной жизни (переписка с иностранцами, транслитерация имен и фамилий для заполнения форм), автоматизированное решение рассматриваемой в работе задачи зачастую не является критическим как для среднего человека, так и для коммерческих компаний.

В случае необходимости английское слово будет транскрибировано носителем языка самостоятельно, хотя и, возможно, не совсем правильно. Однако, другой носитель русского языка скорее всего поймет даже такой вариант, поэтому в повседневной жизни такой точности в большинстве случаев хватает. Онлайн-переводчики (Yandex Translate, Google Translate) обычно вовсе не переводят англоязычные слова и неологизмы (поскольку перевода для таких слов нередко не существует), оставляя их в оригинальном написании. Также нередко они являются именами собственными (так, в [12] приводится цифра в 56-72%), в частности, названиями компа-

ний или организаций. Исключение составляют имена и фамилии людей – они на русский почти всегда переводятся.

Узкая направленность задачи ведет к малому числу предлагаемых решений.

Большинство крупных коммерческих сервисов распознавания речи не нуждаются в решении данной проблемы, так как используют мультязычные модели. Таким образом, данные сервисы не имеют собственного алгоритма решения задачи.

Все найденные авторами решения реализованы как веб-сайты без какого-либо API. Подобные сайты совершенно не предназначены для автоматизированного использования или исследовательской деятельности и нацелены на потребности простых людей. Алгоритм, использованный данными сервисами, также нигде не раскрывается. Однако, исходя из примеров использования, можно сделать вывод о том, что используется подход, основанный на правилах и базе созданных вручную транскрипций для популярных слов.

Одним из наиболее качественных примеров можно выделить сервис на сайте «Английский для занятых» [30], так как, в целом, для существующих в языке слов качество транскрипций вполне при-

емлемое. В частности, реализовано исправление ошибок на основе отзывов пользователей. Однако, правильный вариант для исправленных слов, по-видимому, добавляется в базу вручную и не влияет на правила, породившие неправильный вариант. Поэтому другие, менее употребимые слова, по морфологии похожие на исправленные, остаются без изменений и транскрибируются не совсем правильно. Однако, для работы с OOV данное решение совершенно не подходит: незнакомые сервису слова переводятся по буквам, игнорируя все правила английского произношения: например, “californication” -> “калифорникатион”.

Поиск посвященных теме построения русской транскрипции английских слов

научных статей на русском и английском языках также не дал результатов. Это можно объяснить нишевостью данной задачи.

4 Предлагаемый метод

Для начала отметим, что абсолютно точное решение задачи построения русской транскрипции невозможно в силу различия наборов звуков, соответствующих различным языкам.

Так, звук, соответствующий букве «W» в английском языке, отсутствует в русском языке. Для русскоязычной записи английских слов, содержащих эту букву, согласно принятым правилам, следует использовать различные, в зависимости от следующих букв, двухбуквенные сочетания, первой из которых будет буква «У», например, «Wales» -> «Уэльс». Однако, реальное звучание этих двух слов сильно отличается. Нередко это также зависит от конкретного человека, который будет произносить на месте «W» скорее «В», особенно для тех слов, где официально принятого русского аналога не существует.

Например, для обозначения операционной системы Windows вопреки правилам транскрипции и транслитерации обычно используются слова «Виндовс», «Виндоус», «Виндус», а жаргонное название – «винда». Это свидетельствует о том, что официально принятая система может не отражать ситуации в разговорном языке, и потому не является безусловно верной.

Вольность произношения иноязычных звуков в разговорной речи также сочетается с фактом, что поскольку носитель русского языка зачастую не является носителем английского, он может произносить не те звуки, которые должны быть произнесены по прави-

лам английского, даже если эти звуки имеются в русской речи. Так, в русской речи встречается прочтение написанных на латинице слов, в том числе английских, по буквам, «в стиле» латыни. Например, «data» -> «дата», «Microsoft» -> «Микрософт». Такие транскрипции англоязычных слов будем называть наивными.

Подобные «неправильные» варианты при большой популярности могут закрепляться в речи, что приводит к довольно большому числу более или менее употребимых вариантов произношения одного и того же слова. Так, в русскоязычной литературе, Dr. Watson, персонаж из книг Конан Дойля про Шерлока Холмса, может в разных переводах быть назван как «доктор Ватсон», «доктор Вотсон» и «доктор Уотсон».

Таким образом, это позволяет выдвинуть гипотезу, что включение сразу нескольких вариантов в словарь может повысить вероятность успешной детекции. Это основано на наблюдениях, как сделанных автором, так и сторонних [29], что добавление слов в словарь в большинстве случаев не приводит к ухудшению качества на других доменах и для текстов, не содержащих эти слова.

Общий план алгоритма составления примерных русских транскрипций можно описать следующим образом.

Для каждого английского слова существует транскрипция. Ее можно просмотреть, например, в словаре, в том числе крупных интернет-словарях, например, в Cambridge Dictionary. Для написания транскрипций обычно используется система International Phonetic Alphabet (IPA). Эта система основана на сопоставлении своего символа для каждого звука (в отличие от букв, которые могут обозначать разные звуки в различных ситуациях использования). Для английского языка также была специально разработана система записи ArpaBet, которая согласуется с IPA, но при этом имеет более удобную запись для компьютерного отображения.

Поэтому для получения корректного английского звучания возможно воспользоваться транскрипцией в одной из этих систем. В качестве таких систем можно привести CmuDict, Logios и другие. CmuDict [23] является классическим словарем. Он способен переводить знакомые ему слова в транскрипции в системе ArpaBet. Существует реализация для языка Python в виде библиотеки pronouncing [17]. Его недостатком является невозможность работы со словами, отсутствующими в словаре. Также было замечено, что нередко предлагаемая CmuDict транскрипция отлична от тран-

скрипций, приводимых в авторитетных словарях, таких как Cambridge Dictionary.

Этот недостаток можно компенсировать использованием Logios [13]. Этот сервис позволяет составить транскрипции в системе AgraBet для любых слов согласно правилам английского языка. Таким образом, Logios позволяет получать транскрипции для тех слов, для которых не указана принятая транскрипция в CmuDict. Это особенно актуально в связи с тем, что многие OOV слова могут являться терминами или неологизмами и отсутствовать в любых словарях. Также Logios способен распознавать аббревиатуры. Недостатком Logios является его основанный на правилах подход, который может быть недостаточно гибким в случае нестандартных слов (например, записанные латиницей фамилии).

В качестве аналога Logios можно также использовать g2pE [5], модель на основе алгоритмов машинного обучения. Однако, в процессе экспериментов оказалось, что качество ее транскрипций значительно ниже (что заметно даже визуально: «GitOps (реально читается как «гитопс»))» -> «JH AY T P OW Z», что примерно соответствует произношению «джайтпоуз») и не дает модели воз-

возможности распознать транскрибированные слова, поэтому от использования данной модели было решено отказаться.

Далее, для каждого символа AgraBet предоставляется набор русских букв, которые наиболее точно описывают этот звук. Особо это актуально для гласных звуков. Так, звуку, обозначенному в AgraBet как «АН» могут соответствовать различные гласные – «Э», «А» и «О», поскольку оригинальный английский звук является чем-то средним между звуками, производимыми этими буквами русского алфавита. Использование нескольких вариантов необходимо для того, чтобы учесть различные варианты произношения различными людьми.

Таким образом, для каждого англоязычного слова потенциально может быть построено несколько вариантов русских транскрипций, по одному на каждый вариант произношения каждого из многозначных звуков. Такие слова-транскрипции, соответствующие одному англоязычному слову, будем далее называть гипотезами.

Эксперименты, проведенные автором, показывают, что количество ошибок, связанных с добавлением в словарь лишних слов, значительно меньше, чем количество ошибок, вызванных отсутствием правильных слов в словаре. Поэтому добавление нескольких гипо-

тез обычно не должно приводить к ухудшению качества распознавания. Число гипотез может потенциально очень большим в случае очень длинных слов, поэтому ограничим k (k – параметр алгоритма) гипотезами максимальное число гипотез, которые мы будем строить для одного слова. В случае наличия большего числа гипотез для одного слова, будем брать первые k в порядке построения (что соответствует наиболее частотным вариантам произношения отдельных звуков). Предлагается взять значения k , равные 1 (т.е. выбирается наиболее «вероятная» гипотеза), 2 (этого должно хватать, чтобы охватить все гипотезы для значительной группы слов, при этом не сильно «нагружая» словарь), а также 50 (выбор именно такой «максимальной» константы обусловлен ожиданием в большинстве слов не более 4-5 «спорных» звуков. Для большинства реальных слов достижение этого потолка маловероятно, а на основе проведенных экспериментов можно наблюдать, что для большинства реальных OOV слов получается не более 2 гипотез, что и обуславливает выбор именно этой константы для метрики.

Также дополнительно был реализован алгоритм построения гипотезы на основе наивной транскрипции слова (data -> дата). Обоснование возможности появления такого произношения в речи бы-

ло дано ранее в данной главе. Наивные гипотезы также добавляются в итоговый словарь, если $k > 1$.

Таким образом, алгоритм делится на три основные стадии:

- английское слово сначала переводится в английскую транскрипцию в системе AgraBet с помощью программных средств. Было реализована возможность использования двух различных методов для получения транскрипций незнакомых CmuDict слов: с помощью CmuDict и с помощью, но g2pE показала недостаточный для детекции результат и от ее реального использования было решено отказаться;
- затем транскрипция в AgraBet переводится в соответствующее ей русское слово на основе заранее созданного путем словаря соответствия звуков. Словарь был составлен эмпирическим путем на основе установления примерного отношения между фонемами AgraBet и буквами русского алфавита (напомним, что точное соответствие для большинства звуков невозможно);
- затем, если $k > 1$, к гипотезам, полученным в прошлом пункте, добавляются наивные гипотезы.

Алгоритм был реализован на языке Python3 с применением библиотек `pronouncing` (для работы с `CmuDict`), `requests` (для работы с `Logios`) и `g2p_en` (для работы с `g2pE`).

В качестве подходов к улучшению качества работы алгоритма можно рассмотреть следующие основные направления:

- нахождение более точных баз транскрипций для словарных слов, сравнение полученных транскрипций с вариантами, предлагаемыми авторитетными словарями английского языка;
- рассмотрение более гибких систем составления английских транскрипций для незнакомых слов, основанных на методах машинного обучения;
- уменьшения количества различных вариантов русских букв для некоторых многозначных фонем `АграBet` на основе анализа закономерностей русского произношения и установления более точных соответствий.

5 Постановка экспериментов

5.1 Окружение для проведения экспериментов

Для тестирования описанного алгоритма была использована ранее самостоятельно реализованная система, являющаяся надстройкой над русскоязычной моделью VOSK 0.22. Среди HMM моделей решение от VOSK – одно из лучших по качеству. Модель является вполне конкурентоспособной на рынке русскоязычных ASR моделей, не уступая, а нередко и превосходя по качеству многие современные ей решения от крупных компаний, таких как ЦРТ, Google, Microsoft, Amazon, и даже старые версии модели Whisper, что подтверждается информацией на сайте проекта [24] и исследованиях: как в проведенных автором, так и в сторонних [1], [28]. Также преимуществами VOSK являются открытый исходный код, возможность локального запуска и невысокие требования к ресурсам компьютера, что может стать преимуществом в некоторых сценариях.

Однако, на модель значительно уступает по качеству более современным решениям вроде Silero и новым версиям моделям Whisper. Заметим, что для задачи тестирования и применения может ис-

пользоваться любая другая система, поддерживающая функционал пользовательских словарей вне зависимости от ее архитектуры; предлагаемый алгоритм способен повысить (в большей или меньшей степени в зависимости от самой системы) качество работы любой такой системы и не заточен исключительно под использование с VOSK.

Словарь для данной модели представляет собой список слов, которые модель должна научиться распознавать, с транскрипциями в специально заданном формате. Для построения таких транскрипций система пользуется решением, описанным в документации VOSK, основанном на использовании *Phonetisaurus* [6] (доступном через одноименную библиотеку для Python 3) для построения транскрипций слов словаря, записанных на русском языке. Полученные таким образом слова добавляются в собственный словарь модели. Словарь модели доступен для просмотра в явном виде (т.е. известен список уже известных модели слов). Если слово модели незнакомо, оно гарантированно не может быть распознано.

5.2 Подготовка данных

Поскольку задача не является стандартной, и для проверки предложенного метода требуется аудио особого содержания (русская речь с высоким содержанием слов английского языка, потенциально не имеющих аналогов в русском языке), стандартные датасеты не подходят для проведения экспериментов.

Поэтому было решено использовать выдержки из собственного датасета, подготовленного ранее. Исходный датасет состоит из 176 видео youtube-канала Highload [27] за промежуток с 04.10.2021 по 24.11.2022 различной длины и тематики (в рамках IT индустрии), в том числе записи конференций и интервью с специалистами. Выбор был сделан из следующих соображений: 1) открытый доступ; 2) видео на русском языке; 3) предметная область IT обеспечивает относительно высокое содержание английских слов; 4) формат и качество аудио соответствуют условиям реального применения; 5) разнообразие докладчиков сглаживает индивидуальные особенности произношения.

Для выбранных видео с помощью библиотек `yt_dlp` [7] и `ffmpeg` [4] для языка программирования Python 3 были сделаны отрезки, содержащие первые 5 минут видео. Видео приходилось размечать

вручную, поскольку техническая сложность аудио и особенности разметки для англоязычных слов делает неэффективным использование сервисов вроде Яндекс.Толоки, что приводит к очень высоким временным затратам. Для каждого отрезка C_i был также составлен список W_i встреченных англоязычных слов (не более 10, если в данном отрезке было больше, брались случайные 10, если меньше – брались все встреченные), не имеющих транскрибированных аналогов в словаре модели. Для каждого англоязычного слова w_j из W_i с помощью описанного в прошлой главе алгоритма было построено множество всех русскоязычных слов-гипотез $H_{ij} = \{h_{ij1}, \dots, h_{ijx}\}$, где $x \leq k$ – параметр алгоритма. Итоговый словарь V_i представляет собой объединение всех H_{ij} для одного отрезка видео (C_i). Нетрудно посчитать, что итоговый размер словаря для одного видео не превышает $10 * k$ слов, что для исследуемых значений параметра k дает максимальный размер от 10 до 500 слов.

5.3 Формулировка эксперимента

Было проведено два типа экспериментов.

В экспериментах первого типа дообученная на словаре модель V затем применялась к тому же видео, из которого были извлечены слова w_i , использованные для составления словаря V . Этот механизм позволяет гарантировать наличие искомых слов в аудио и оценить реальный результат алгоритма транскрибирования в отрыве от других компонентов системы. Однако, в реальной жизни мы не можем знать, какие именно слова встретятся в видео. Эта проблема является темой отдельных исследований, выходящих за рамки данной работы.

Сильно упрощенная модель «слепых» экспериментов, где словарь строится без опоры на тестируемое видео, была реализована в экспериментах второго типа. В ходе такого эксперимента для англоязычных слов w_j из исходного видео C_i создается запрос Q_i во внутреннем поиске youtube вида « $w_1 w_2 \dots w_{10}$ русский», т.е. длина $\text{len}(Q_i) = \text{len}(W_i) + 1$. По запросу Q_i были извлечены несколько первых видео на русском языке, на 5-минутных отрезках которых проходит тестирование. Обозначим эти отрезки как $C_{i,1}$, $C_{i,2}$, и т.д. Добавление слова «русский» в конец запроса Q_i необходимо для того, чтобы результаты выдачи были на русском языке (поскольку слова w_j – англоязычные, то результаты по запросу « $w_1 w_2 \dots w_{10}$ »

нередко оказываются преимущественно на английском языке), так как встроенный фильтр Youtube не позволяет фильтровать выдачу по языку видео.

Предполагается, что выдача по такому запросу будет содержать преимущественно русскоязычные видео, которые с высоким шансом будут содержать указанные в запросе слова. А значит, качество распознавания отрезков $C_{i,j}$, составленных из этих видео, при использовании словаря V_i , составленного для исходного видео C_i , должно быть немного выше, чем без него. Однако, на практике, из-за особенностей поиска youtube, поиск практически игнорирует не частотные слова, что приводит к тому, что искомые слова появляются в найденных видео в очень небольших количествах.

5.4 Используемая метрика качества

В качестве метрики оценки качества используется модифицированная версия стандартной для таких задач метрики WER [25]; чем меньше значение метрики, тем выше качество.

Модификация предназначена для уменьшения шума, вызванного ошибками в незначущих словах и окончаниях, и заключается в автоматизированной предобработке текстов – результата модели и

эталонного текста. Предобработка включает в себя исключение небуквенных (знаки препинания) и излишних пробельных символов, стемминг, замену чисел на числительные и удаление незначащих слов – некоторых коротких союзов и предлогов. Это позволяет увеличить вес значащих ошибок, в том числе, ошибок в OOV словах. За счет устранения шума во флективной лексике метрика становится более чувствительна и отображает фактические изменения даже при разнице в 1-2 по-другому распознанных слова.

6 Результаты экспериментов

6.1 Результаты

Рассматривались следующие видео файлы.

Видео C_{77} : <https://www.youtube.com/watch?v=F0Zif5Enbgc>

Список OOV слов W_{77} : {microburst, incast, TCP}.

Словарь V_{77} для $k=2$: {майкроубёрст, мицробурст, инкэст, инцаст, тисипи, тцп}.

Таблица 1. Результаты экспериментов первого типа. В левом столбце таблицы записано название видео в виде C_i , где i – номер видео в исходном датасете, url-адрес видео написан под таблицей. Значение $\text{len}(W_i)$ для видео указано во втором столбце. В остальных столбцах указано значение WER для соответствующего видео, выраженное в процентных пунктах ($\text{WER} * 100 \%$) в зависимости от использования словаря V_i и параметра k максимальной длины словаря.

Название видео	Число ан-глийских слов W_i	WER, без слов-варя	WER, $V_i, k=1$	WER, $V_i, k=2$	WER, $V_i, k=50$
C_{144}	$\text{len}(W_{144}) = 4$	15.54 %	15.14 %	15.14 %	15.14 %
C_{77}	$\text{len}(W_{77}) = 3$	6.88 %	6.38 %	6.38 %	6.38 %
C_{162}	$\text{len}(W_{144}) = 4$	11.66 %	11.66 %	11.66 %	11.66 %
C_{29}	$\text{len}(W_{77}) = 3$	14.90 %	14.71 %	14.71 %	14.71 %

Видео C_{144} : <https://www.youtube.com/watch?v=irlVDHvTAEM>

Список OOV слов W_{144} : {integrity, devops, PM, highload}.

Словарь V_{144} для $k=2$: {интегрэти, интегрити, дэвопс, девопс, пи-ем, пм, хайлоуд, хигхлоад}.

Видео C_{162} : https://www.youtube.com/watch?v=5HBp0vciV_A

Список OOV слов W_{162} : {IPO, CTO, MVP, CPO}

Словарь V_{162} для $k=2$: {ситио, цто, айпио, ипо, емвипи, мвп, сипио, цпо}

Видео C_{29} : <https://www.youtube.com/watch?v=zcleMmzoeCo>

Список OOV слов W_{29} : {HTTP, latency, TLS}

Словарь V_{29} для $k=2$: {эйчтитипи, хттп, лэйтэнси, лэйтанси, тиелес, тлс}

Таблица 2. Результаты экспериментов второго типа. В левом столбце таблицы записано название видео, url-адрес видео написан под таблицей. Словарь V_i построен по видео с названием C_i из предыдущей таблицы, запрос Q_i содержал $\text{len}(Q_i)$ слов. Значение $\text{len}(Q_i)$ для видео указано во втором столбце. Полученные по запросу Q_i видео именуются как $C_{i,j}$, где $j=1,2,3$ и т.д. – порядковый номер в выдаче. В остальных столбцах указано значение WER для соответствующего видео, выраженное в процентных пунктах ($\text{WER} * 100\%$) в зависимости от использования словаря V_i и параметра k максимальной длины словаря.

	длина	за-	WER,	WER,	WER, V_i	WER, V_i
	проса Q_i	без	сло-	$V_i, k=1$	$k=2$	$k=50$
		варя				
$C_{144,1}$	$\text{len}(Q_{144}) = 5$	13.09 %	12.47 %	12.47 %	12.47 %	12.47 %
$C_{144,2}$	$\text{len}(Q_{144}) = 5$	22.77 %	22.94 %	22.94 %	22.94 %	23.08 %
$C_{77,1}$	$\text{len}(Q_{77}) = 4$	7.87 %	7.87 %	7.87 %	7.87 %	7.87 %

Запрос Q144: {integrity, devops, PM, highload, русский}.

Запрос Q77: {incast, TCP, microburst, русский}.

Видео C_{144,1}: <https://www.youtube.com/watch?v=v0-d5Z9C9u4>.

Видео C_{144,2}: <https://www.youtube.com/watch?v=5PcmBdfGpOo>.

Видео C_{77,1}: <https://www.youtube.com/watch?v=oRtYn2KZFRY>.

6.2 Анализ

На основании полученных результатов можно сделать следующие наблюдения:

Во-первых, при использовании словаря частота детекции незнакомых англоязычных слов в большей части экспериментов увеличивается (становится ненулевой). Однако, улучшение WER при текущей постановке эксперимента довольно незначительно. Это объясняется следующими причинами:

- несмотря на предпринятые усилия, сформированный датасет оказался недостаточно удачным для рассматриваемой задачи, так как для многих видео доля незнакомых английских терминов оказалась недостаточно велика;

- слова, попадающие в словарь, нередко не являются достаточно частотными, и потому практически не влияют на WER даже при корректном распознавании;
- в экспериментах второго типа не все из заданных в запросе Q слов реально встречаются в аудио, особенно это актуально для имен собственных;
- очень плохо распознаются даже корректно транскрибированные аббревиатуры. Это происходит в том числе из-за ошибок при определении ударений, осуществляемом Phonetisaurus, который встроен в пакет обновления модели VOSK. Так, из-за неправильного ударения не распознавалось слово «тисипи», являющееся верной транскрипцией для OOV слова «TCP». Для повышения вероятности корректного распознавания моделью аббревиатур может использоваться автоматизированное исправление ударений в транскрипциях, генерируемых Phonetisaurus, на основе факта, что обычно в аббревиатурах ударение падает на последний слог;
- из-за высокой флективности русского языка, выражающейся в обилии словоформ и словообразования, может не происходить распознавание даже для корректно транскрибированных OOV

слов (например, не распозналось слово «девопсеры», когда запрос Q содержал слово «devops» с верной русской транскрипцией «девопс»);

- при отсутствии слов из словаря в тексте качество распознавания остальных слов обычно не изменяется;
- падение качества при использовании словаря происходит из-за появления дополнительных ошибок вставки, вызванных созвучием.

Также можно заметить, что значение параметра k практически не влияет на качество. Это может быть связано с тем, что получаемые слова достаточно созвучны, и модель распознает их приблизительно одинаково. В связи с этим перспективным полагается использование $k=2$, поскольку это минимальное k , при котором в словарь попадают как обычные транскрипции, так и «наивные», которые в некоторых случаях могут значительно отличаться от эталонного произношения.

Для получения более точной оценки численной эффективности алгоритма необходима разработка метода подготовки специализированного датасета с более высокой долей частотности англо-

язычных терминов и пересечений по таким терминам между разными аудио.

7 Заключение

В данной работе были рассмотрены методы повышения качества распознавания русской речи, содержащий англоязычные термины. Наличие таких терминов характерно для многих предметных областей, например, для области ИТ, и потому их корректное распознавание является актуальной задачей.

Предложен и реализован алгоритм составления словаря, состоящего из т.н. русских транскрипций для английских слов, которые предлагается формировать путем сопоставления букв русского языка с символами AgraBet, содержащимися в англоязычной транскрипции соответствующих слов. Применение словаря должно позволить повысить качество работы для любой ASR системы, поддерживающей функционал расширения словаря.

Полученные в результате экспериментов с системой на основе модели VOSK данные свидетельствуют о практической эффективности предложенного алгоритма. Качество распознавания в среднем повышается для результатов распознавания видео, потенциально

релевантных словарю и предметной области. С другой стороны, качество практически не падает при использовании словаря при работе с нерелевантными видео. Это позволяет говорить о возможности применения описанного алгоритма для улучшения качества работы ASR систем.

В качестве дальнейших направлений исследований предлагается улучшение отдельных составляющих алгоритма, опирающихся на внешние системы транскрибирования, а также составление более подходящего для данной задачи специализированного датасета для тестирования алгоритма.

8 Благодарности

Исследование выполнено в рамках государственного задания МГУ имени М.В. Ломоносова.

9 Соблюдение этических норм сенатором

Данная статья является полностью оригинальным научным произведением ее авторов, прежде не опубликована и до получения решения редколлегии PRIA о непринятии ее к публикации не будет направляться в другие издания.

10 Раскрытие потенциальных конфликтов интересов

Процесс написания и содержание статьи не дает оснований для постановки вопроса о конфликте интересов.

11 Литература

1. A. Kuznetsov, Сравнение Vosk и Whisper. <https://habr.com/ru/articles/814057/>. Accessed May 30, 2024.
2. A. Veysov, Ультимативное сравнение систем распознавания речи: Ashmanov, Google, Sber, Silero, Tinkoff, Yandex, <https://habr.com/ru/post/559640>. Accessed May 30, 2024.
3. Amazon Transcribe Documentation, <https://docs.aws.amazon.com/transcribe>. Accessed May 30, 2024 (VPN required).
4. Ffmpeg - A complete, cross-platform solution to record, convert and stream audio and video, <https://ffmpeg.org/>. Accessed May 30, 2024.
5. g2pE: A Simple Python Module for English Grapheme To Phoneme Conversion, <https://github.com/Kyubyong/g2p>. Accessed May 30, 2024.
6. GitHub – Phonetisaurus g2p, <https://github.com/AdolfVonKleist/Phonetisaurus>. Accessed May 30, 2024.
7. GitHub - yt-dlp: A feature-rich command-line audio_video downloader, <https://github.com/yt-dlp/yt-dlp>. Accessed May 30, 2024.
8. Gladia - OpenAI Whisper vs Google Speech-to-Text vs Amazon Transcribe: The ASR Rundown, <https://www.gladia.io/blog/openai-whisper-vs-google-speech-to-text-vs-amazon-transcribe>. Accessed May 30, 2024.
9. Google Cloud Speech-to-Text, <https://cloud.google.com/speech-to-text/docs>. Accessed May 30, 2024.
10. H. Wang et al., “Improving pre-trained multilingual models with vocabulary expansion”, arXiv preprint arXiv:1909.12440 (2019). <https://doi.org/10.48550/arXiv.1909.12440>
11. I. Sheikh et al, ”Modelling semantic context of OOV words in large vocabulary continuous speech recognition”. //IEEE/ACM Transactions on Audio, Speech, and Language Processing. – T. 25. – №. 3. – С. 598-610 (2017).
12. I. Sheikh et al., “OOV proper name retrieval using topic and lexical context models”, 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2015. – С. 5291-5295 (2015).
13. LOGIOS Lexicon Generation Tool, <http://www.speech.cs.cmu.edu/tools/lextool.html>. Accessed May 30, 2024.
14. N. Kitaoka, B. Chen, Y. Obashi, “Dynamic out-of-vocabulary word registration to language model for speech recognition”, EURASIP Journal on Audio, Speech, and Music Processing 2021:4 (2021).
15. openai_whisper: Robust Speech Recognition via Large-Scale Weak Supervision, <https://github.com/openai/whisper>. Accessed May 30, 2024.
16. P. Fetter. Detection and transcription of OOV words, Ph. D. Thesis (Technical University of Berlin, Germany, 1998).
17. pronouncing – PyPI, <https://pypi.org/project/pronouncing/>. Accessed May 30, 2024.

18. R. Prabhavalkar et al., “E2E speech recognition: A survey”, IEEE/ACM Transactions on Audio, Speech, and Language Processing (2023). <https://doi.org/10.48550/arXiv.2303.03329>.
19. S. Ashmanov, Как мы проверили качество распознавания речи у Яндекс, Гугла, Тинькофф, Amazon и др., https://habr.com/ru/company/ashmanov_net/blog/576612/. Accessed May 30, 2024.
20. S. Wang et al, “Now it sounds like you: Learning personalized vocabulary on device”, arXiv preprint arXiv:2305.03584 (2023). <https://doi.org/10.48550/arXiv.2305.03584>.
21. Silero Speech-To-Text, <https://audio-v-text.silero.ai/>. Accessed May 30, 2024.
22. Speech to Text – Audio to Text Translation | Microsoft Azure, <https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/#overview>. Accessed May 30, 2024.
23. The CMU Pronouncing Dictionary, <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>. Accessed May 30, 2024.
24. VOSK Offline Speech Recognition API, <https://alphacephei.com/vosk>. Accessed May 30, 2024.
25. Word error rate, https://en.wikipedia.org/wiki/Word_error_rate. Accessed May 30, 2024.
26. X. Zheng et al., “Using synthetic audio to improve the recognition of out-of-vocabulary words in E2E ASR systems”, ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – IEEE, 2021. – С. 5674-5678 (2021). <https://doi.org/10.1109/ICASSP39728.2021.9414778>
27. YouTube - HighLoad Channel, https://www.youtube.com/channel/UCwHL6WHUarjGfUM_586me8wl. Accessed May 30, 2024 (VPN required).
28. Д. Маненок, “Применение N-граммных языковых моделей для коррекции доменных коллокаций и улучшения распознавания речи”, Научный аспект - Информ. Технологии 2023; 4 (2023).
29. Как Улучшить Качество Распознавания Речи? <https://www.silero.ai/how-to-improve-stt-quality>. Accessed May 30, 2024.
30. Транскрипция английских слов русскими буквами: Английский для занятых, <https://englishforbusy.ru/transliterator/>. Accessed May 30, 2024.

12 Краткие биографии и фотографии всех авторов



Волков Евгений Владимирович, 1997 г.р., аспирант кафедры алгоритмических языков факультета ВМиК МГУ имени М.В. Ломоносова, работает в Научно-исследовательском вычислительном центре МГУ. Область научных интересов – распо-

знавание речи, особенно в специальных предметных областях.



Добров Борис Викторович, 1963 г.р., кандидат-физико-математических наук, заведующий лабораторией Научно-исследовательского вычислительного центра МГУ. Окончил факультет ВМиК МГУ имени М.В. Ломоносова в 1985 г. Область научных интересов – обработка и структурирование больших потоков неструктурированных документов, формирование больших лингвистических онтологий и графов знаний, интеграция онтологических и нейросетевых подходов обработки текстов, построение информационно-аналитических систем.