Meaning Error Rate

Людмила Гордеева
Университет ИТМО
lulu.gordeeva07@gmail.com

Василий Ершов
Яндекс
noxoomo@yandex-team.ru

Игорь Лабутин
Яндекс
НИУ ВШЭ
Labutin.IgorL@gmail.com

Игорь Кураленок
Яндекс
solar@yandex-team.ru

***Аннотация*—Стандартный подход для оценки качества систем распознавания речи основан на вычислении числа неправильно распознанных слов (метрика WER, Word Error Rate). Данный подход является простым и достаточно эффективным в большинстве случаев. Однако, он не учитывает несколько существенных факторов. Во-первых, системы распознавания голоса могут допускать ошибки, существенно меняющие смысл фразы, а метрика WER не является чувствительной к данному типу ошибок. Во-вторых, в реальных приложениях от систем распознавания голоса требуется не точная транскрибация слово в слово, а специфичная для конкретного приложения функциональность: возможность определить потребность пользователя (интент) для голосовых ассистентов; качество распознавания автомобильного номера, адреса, номера телефона и т.д. Таким образом, возникает необходимость в разработке метрик для оценки качества распознавания в конкретных приложениях, а также необходимо научиться оценивать качество современных систем не только в терминах числа ошибочно распознанных слов, но и с учетом того, что целью является передача смысла фразы. В данной работе мы разработали общий подход к построению такого типа метрик, а также способ оценки качества метрик. Основная идея нового подхода — использование краудсорсинга и последующего сведения задачи построения метрики к хорошо изученной задаче обучения с учителем. В качестве примера использования данного подхода мы предлагаем обобщение метрики WER — метрику MERaLM, обладающую следующими**

достоинствами: учет того, что не все ошибки одинаково влияют на точность передачи смысла фразы и легкая интерпретируемость.

Ключевые слова— автоматическое распознавание речи, метрика, машинное обучение.

I. Введение

Автоматическое распознавание речи в настоящее время активно применяется для решения большого числа практических задач: голосовые помощники, автоматическая генерация субтитров, автоматизация работы центров обработки звонков и т.д. Современные методы, основанные на применении методов глубинного обучения, достигли качества, сравнимого с человеком [13] на общедоступных корпусах транскрибированных аудио таких, как LibriSpeech. В практических задачах требуется работать с шумными данными, часто плохого качества (например, записи звонков в телефонии), для которых разработанные системы показывают существенно худшее качество. Тем не менее, этого качества уже достаточно для того, чтобы активно внедрять автоматическое распознавание в различные приложения. Большое количество работ в данной области направлено на то, чтобы улучшать качество работы систем распознавания для различных практических ситуаций. К сожалению, основной фокус в данных работах направлен на эксперименты с моделью распознавания, а именно эксперименты с архитектурой и типом нейронной сети (использование сверточных [11; 14] или рекуррентных сетей [8], архитектуры на основе трансформеров [16], и т.д.), методами аугментации и другими техниками обучения глубинных сетей, а также способами ускорить обучение и применение систем распознавания в реальных задачах. В то же время, все эти методы сравниваются на основе единственной метрики — числе ошибочно распознанных слов (WER), которая хотя и сильно коррелирует с качеством распознавания, тем не менее не является чувствительной к ошибкам, существенно меняющим суть произнесенного текста. Таким образом, улучшения моделей по существующей метрике не позволяют оценить влияние тех или иных изменений на то, как эти изменения будет воспринимать пользователь. В результате, в распознавании речи сегодня возникает проблема, близкая к той, с которой столкнулся информационный поиск в 2000-х годах, когда стало понятно, что не все документы одинаково хорошо отвечают на запрос пользователя, а существующие метрики, такие, как MAP, не позволяют улучшать системы в сторону выявления более релевантных документов [5]. В качестве решения было предложено семейство метрик DCG и NDCG [5]. В распознавании речи сегодня наблюдается аналогичная проблема — отсутствует подход к построению метрик, отражающих то, как пользователи воспринимают работу системы: какое распознавание считать *"*хорошим*"*, а какое *"*плохим*"*? Передает ли распознавание смысл исходного текста? Какие ошибки считать существенными, а какие нет? Насколько влияет добавление или пропуск частицы *"*не*"* на восприятие смысла? Человек, так или иначе, может дать ответы на эти вопросы, в то время как метрика WER — нет, и мы считаем, что для дальнейшего развития систем распознавания голоса нужно понять, что и зачем мы хотим оптимизировать.

Наиболее эффективный подход к построению оценки качества систем распознавания голоса — привлечь к этой задаче человека (например, на основе краудсорсинга), который сможет формализовать такие понятия, как *"*смысл фразы передан точно*"*. Однако экспертная оценка стоит дорого и требует большого количества времени и усилий на проведение любого эксперимента. Поэтому в данной

работе мы предлагаем решение, близкое по духу к тому, которое было применено для оценки ранжирования в информационном поиске с помощью метрики DCG и NDCG; и в системах автоматического перевода при переходе к оцениванию с помощью метрики BLEU [4]. Вместо того, чтобы привлекать экспертов к оценке каждой системы распознавания, мы их используем один раз для обучения модели, которая будет автоматически предсказывать рейтинги экспертов.

II. Обзор существующих метрик

На данный момент в задачах автоматического распознавания речи используется метрика WER. Она является стандартом и используется для оценки качества систем. Однако, несмотря на свою популярность, эта метрика имеет ряд существенных недостатков. Основным из них является одинаковый вес ошибки. Как следствие, ошибки в словах сильно искажающих смысл и ошибки в незначительных для содержания словах будут иметь одинаковые веса.

На основе WER было разработано большое количество метрик, которые также применяются в области автоматического распознавания речи и вот некоторые из них: расстояние Левенштейна [1], WER with embeddings [10], MR-WER (Multi-Reference WER) [9], Match Error Word, Normalized WER, WIL (Word Information Lost)[7].

Эти метрики созданы в качестве альтернативы WER для применения в более узких областях: особенности языка, особенности конкретной задачи и т.д. Однако их основная идея состоит в подсчете количества ошибок. Такая же идея лежит и в основе WER. Таким образом, все эти метрики обладают одинаковыми недостатками.

Стоит отметить, что автоматическое распознавание голоса не единственная область машинного обучения, где используются метрики для оценки качества через сравнение пар предложений. В машинном переводе стандартной метрикой качества является BLEU (Bilingual evaluation understudy) [4]. Но метрики машинного перевода не подходят для задачи распознавания речи по двум причинам. Во-первых, это другая задача, для которой нет одного правильного ответа, в отличии от распознавания (в данной работе не рассматриваются языки, имеющие неоднозначное орфографическое представление). Во-вторых, ни одна метрика, использующаяся в машинном переводе, не отображает отсутствие или наличие семантических различий предложений.

Задача машинного перевода интересна тем, что в ней были предприняты попытки внедрить идею передачи семантической составляющей в метрику. Так появилась метрика NIST [6]. Она является усовершенствованной версией стандартной метрики BLEU, но учитывает то, что штраф за разные ошибки должен быть разным. Величина штрафа обратно пропорциональна встречаемости ошибки. Такой способ позволяет давать малые штрафы за мелкие ошибки в артиклях, а большие за ошибки в значимых словах. При таком способе формирования штрафа хорошо видна проблема того, что замена слова на его синоним, будет *"*стоить дорого*"*, а значит отображение семантических свойств предложений все еще не отображается.

Остальные метрики машинного перевода, являются различными модификациями метрики BLEU и не подлежат рассмотрению в рамках данной статьи.

III. Методология сравнения метрик

Мы выяснили какую задачу должна решать метрика для оценки качества систем распознавания речи, рассмотрели существующие метрики и выявили ряд основных недостатков. Теперь необходимо разработать новый способ оценки, который будет лишен этих проблем.

Первая задача, которая возникает при разработке новой метрики — формализация понятия того, что одна метрика лучше чем другая. Одним из естественных способов такой формализации является определение необходимого набора свойств, которыми должна обладать *"*хорошая*"* метрика. В некоторых приложениях удается доказать, что такой набор свойств определяет метрику однозначно. Так, например, для некоторых моделей машинного обучения используют технику SHAP values [15] из теории игр. Она позволяет, в некотором смысле, оптимально оценить вклад разных признаков в итоговую модель. Аналогичные идеи можно использовать и при построении метрик оценки качества. Тогда критерий качества метрики формализуется так — любая метрика, удовлетворяющая набору свойств (по сути, аксиом). Основная цель — составить такой набор аксиом, что метрика, удовлетворяющая этим аксиомам, единственна. В случае прикладных задач выбор метрики — выбор наиболее *"*хорошего*"* для конкретной задачи набора аксиом.

К сожалению, такой подход в общем случае не реализуем. Во-первых, составление списка аксиом трудная и, во многом, субъективная задача. Во-вторых, даже имея список аксиом, построить метрику, удовлетворяющую им, а также доказать, что такая метрика единственна, выглядит как нерешаемая задача.

Поэтому в данной статье мы предлагаем другой подход. Вместо того, чтобы математически корректно вводить набор аксиом и доказывать, что какие-то метрики им удовлетворяют, мы предлагаем сформулировать некоторую *"*идеальную*"* метрику для оценки качества распознавания речи. Основная цель метрики — согласованность с человеком в рамках конкретной задачи. Поэтому идеальной метрикой можно назвать эксперта и подробную инструкцию, по которой эксперт сможет однозначно сопоставить паре предложений (истинному тексту на аудио и гипотезе) некоторую оценку качества распознавания под требования конкретного приложения. Тогда качество метрики в рамках определенной прикладной задачи — это мера согласованности с такой *"*идеальной*"* метрикой. Ту метрику, которая лучше согласована, будем считать лучшей.

Чтобы применять такой подход для оценки качества метрики в рамках конкретной прикладной задачи необходимо формализовать понятие согласованности метрики

с ”идеальной”. ”Идеальная” метрика обладает одним существенным недостатком — высокая стоимость. Человеческий труд стоит очень дорого, поэтому использовать его непрерывно и в больших масштабах нецелесообразно. Воспользуемся подходом, использующимся в задачах машинного обучения — подготовим репрезентативный набор данных и получим значения ”идеальной” метрики только на нем. Подготовка такого набора — отдельная и трудная задача. В рамках данной статьи нет возможности уделить ей необходимое внимание. Таким образом, чтобы упорядочить метрики качества распознавания, достаточно один раз посчитать для каждой из них меру согласованности с ”идеальной” на фиксированном репрезентативном наборе данных.

Заметим, что в описанном процессе не формализовано понятие согласованности. Оно связано непосредственно с множеством значений метрики. В экспериментах, которые будут описаны в этой статье для простоты выбрана бинарная шкала оценивания. В таком случае в качестве меры согласованности может быть использована любая метрика качества модели бинарной классификации. Для определенности мы будем использовать AUC — area under the curve (одна из основных метрик для оценки качества бинарной классификации [3]).

IV. Подход к построению семейства метрик

Основной принцип построения метрики — согласованность с ”идеальной” метрикой. Согласованность измеряется на некотором фиксированном наборе данных, репрезентативно представляющем конкретную предметную область. При этом, в качестве метрики можно рассматривать любую функцию от пары предложений. Как следствие, новой метрикой может быть любая функция пар предложений, однозначно описывающаяся некоторым набором параметров. Тогда задача построения новой метрики сводится к выбору модели, сбору данных для обучения этой модели и, собственно, обучения.

В качестве базовой модели была выбрана модификация стандартной метрики WER. Как и в метрике для машинного перевода NIST [6] попробуем усовершенствовать функцию штрафа, рассмотрев произвольную функцию от пар слов, характеризующуюся набором параметров.

Формально можно записать ее следующим образом:

$$\text{MERa} = \min \sigma \left(\mathbb{E}_{(w_i,w_j)\sim\hat{D}} Q(w_i, w_j)) \right),$$

где

$\sigma(x) = \frac{e^x}{1+e^x}$ — сигмоидальная функция,

$E_{(w_i,w_j)\sim\hat{D}}$ — эмпирическое математическое ожидание (по парам слов (w_i, w_j)),

$Q(u, v)$ — стоимость замены слова u на слово v.

Минимум берется по всем ”выравниваниям” предложений. Под ”выравниванием” подразумеваем следующее — каждому слову из первого предложения ставим в соответствие слово из второго (что соответствует замене) или пустое слово (что соответствует удалению). Если во втором предложении остались слова без пары, им в соответствие ставится пустое слово (что соответствует операции вставки). Таким образом, пара предложений однозначно представляется в виде пар слов, некоторые из которых могут быть пустыми.

В результате получается семейство метрик MERa (Meaning error rate), где каждый элемент семейства задается своей функцией стоимости.

Далее рассмотрим базовый пример стоимостной функции, который был реализован.

V. Стоимость исправления смысловой ошибки

В качестве функции стоимости можно рассматривать абсолютно любую функцию. Для базовой реализации используем линейную функцию от выбранных признаков для пары слов.

Выбор признаков, а также данные для обучения модели являются ключевыми элементами процесса построения новой метрики. Основная сложность заключается в том, что во многом эти составляющие взаимосвязаны и выбрать признаки, которые наиболее точно будут отображать влияние на итоговый смысл фразы очень сложно без данных, репрезентативно представляющих предметную область. Как уже было ранее отмечено, сбор такого набора данных является отдельной крайне трудоемкой задачей и выходит за рамки этой статьи. В связи с этим, выбор признаков был основан на некоторых эмпирических предположениях о важности с точки зрения совпадения по смыслу, однако в дальнейшем этот список можно и нужно расширить и возможно найти более значимые. Выбранный набор признаков в таблице I.

Таблица I
Описание признаков

Признак
совпадает словарная форма слова
исходное слово — ”нет”
распознанное слово — ”нет”
слова полностью совпадают
исходное слово — ”не”
распознанное слово — ”не”
исходное слово — ”да”
распознанное слово — ”да”

По сути, функция, которую мы строим, соответствует некоторой модели бинарной классификации (в базовой реализации — линейной). Модель обладает некоторым набором параметров, которые заранее не известны, а значит, их нужно как-то оценить (для линейной — набор коэффициентов при признаках).

VI. Оценка параметров

1) Подготовить обучающее множество. В качестве наблюдений в нашей задаче выступают пары (распознанный текст, исходный текст на аудио), а в качестве целевой переменной — оценка эксперта о

том, насколько точно предсказание передает смысл текста. В наших экспериментах мы использовали бинарные предсказания, но в дальнейшем можно легко добавить вероятностные предсказания.

2) Метод оценки, по сути, функция, которую мы строим, соответствует некоторой модели бинарной классификации (в базовой реализации — линейной). Модель обладает некоторым набором параметров, которые заранее не известны, а значит их нужно как-то оценить (для линейной — набор коэффициентов при признаках).

VI-A. Данные

В качестве набора обучающих данных были собраны результаты распознавания голоса нескольких облачных сервисов компаний Яндекс и Google. Исходные аудио были собраны на основе реального набора данных телефонии (полученные результаты распознавания доступны по ссылке https://github.com/gordeeva-ln/MERa).

Собранный набор данных представляет из себя множество пар — исходный текст аудио и текст, распознанный одним из указанных выше сервисов.

Далее на его основе создали задание для сервиса Яндекс.Толока, в которой попросили людей (экспертов) разметить пары собранного набора данных на два класса. К первому классу относятся предложения, совпадающие по смыслу, ко второму — различающиеся. Заметим, что не для любой пары предложений можно однозначно установить, к какому классу нужно ее отнести. Поэтому каждая пара была отправлена одновременно трем пользователям. Таким образом, был собран набор данных, использующийся для эксперимента и содержащий в себе около 5000 размеченных пар.

VI-B. Оценка параметров

Наша задача свелась к стандартной логистической регрессии. Для оценки оптимальных параметров необходимо найти минимум логистической функции потерь. Сложность задачи заключается в том, что стандартные методы оптимизации применять не получается, т.к. оптимальное выравнивание зависит от параметров, которые мы оцениваем, а методы непрерывной оптимизации можно использовать только при фиксированном выравнивании. Поэтому мы используем некоторый аналог *EM*-алгоритма [2], на каждой итерации сначала фиксируем оптимальное выравнивание при текущей оценке параметров, затем обновляем оценку параметров с помощью градиентного спуска, затем снова пересчитываем оптимальное выравнивание.

VI-C. Представление данных

Чтобы пояснить дальнейшие рассуждения, введем несколько вспомогательных обозначений:

- Один элемент обучающего множества — пара предложений.
- В процессе подсчета метрики MERa для каждой пары предложений получаем набор пар слов. Метрика для пары предложений — среднее значение Q на парах слов — поэтому порядок слов не важен (важен лишь набор пар).
- Составляем список уникальных пар по всему набору данных и представляем каждый элемент исходных данных как вектор, где каждый элемент равен количеству раз, которое соответствующая уникальная пара встречается в множестве пар для этой пары предложений (матрица C).
- Таким образом, получаем равенство $\mathbf{y_{pred}} = CS$.
  Такое представление данных удобно в контексте используемого алгоритма.

Таблица II
Используемые обозначения

Обозначение	Описание
X	исходные данные (набор пар предложений)
$\mathbf{y_{train}}$	ответы асессоров (0 если смысл одинаковый, 1 если разный)
$\mathbf{y_{pred}}$	значение метрики для пары предложений
Q	функция стоимости для пары слов
U	матрица признаков для уникальных (внутри датасета) пар
S	вектор значений score на соответствующих уникальных парах
C	матрица, в которой элемент (i, j) соответствует количеству раз, которое уникальная пара под номером j встречается в паре предложений под номером i

VI-D. Алгоритм

1) Выбираем некоторую начальную модель, предсказывающую метрику для пары слов. В случае модели, являющейся линейной функцией, достаточно задать набор стартовых весов (можно выбирать случайно).
2) Разбиение на пары:
   Согласно текущей модели, для каждой пары предложений получаем оптимальное (минимально возможное) значение метрики при наилучшем возможном "выравнивании" (разбиении на пары слов) и соответствующее разбиение. Затем получаем матрицы C, S и U. Оптимальное выравнивание ищется с помощью динамического программирования, аналогичного динамическому программированию для вычисления WER, с той лишь разницей, что вес удаления, вставки и замены одного слова на другое вычисляется с помощью модели, а не является константным.
3) Шаг градиентного спуска:
   Посчитать производную функции потерь непосредственно по параметрам модели может быть достаточно сложно, поэтому возьмем производную по S. При помощи представления $y_{pred} = CS$ производную по S посчитать легко.
   Обучим новую модель, которая по U (уникальным парам) будет предсказывать направление градиентного спуска по S и добавим ее к существующей.

4) Повторяем шаги 2-3 пока алгоритм не начнет переобучаться.

VII. Эксперименты

На основе предложенной модели и собранных данных был проведен набор экспериментов. Полученные коэффициенты при выбранных параметрах линейной модели в таблице[1].

Анализируя полученные коэффициенты, можно сделать несколько выводов. Во-первых, в основном значения очень малы. Такого результата можно было ожидать, так как выбор признаков был сделан не исходя из экспериментов, а исходя из интуитивных представлений. Вторая причина в том, что в качестве данных были выбраны реальные распознавания голоса. Ошибки, с которыми не справляются существующие на данный момент метрики и которые при этом можно описать в виде признаков от пар слов, встречаются крайне редко, но именно на такие ошибки метрика должна реагировать. Поэтому так важно и для обучения, и для сравнения метрик между собой собрать подробный репрезентативный набор данных. В рамках данной работы, не было задачи показать насколько сильное влияние оказывают выбранные признаки. Во-вторых, наибольший положительный вес имеет признак, отвечающий за полное совпадение слов. По полученным выводам нельзя судить о значимости признаков, они всего лишь отражают данные, на которых модель обучалась.

Стоит отметить, что до сих пор про признаки было сказано только то, что корректно выбрать их можно после тщательного анализа данных, но как этот анализ проводить и как выделять признаки — отдельная задача.

Теперь переходим к оценке и попытаемся предсказать для пары предложений эквивалентны они по смыслу или нет. Для текущего набора коэффициентов результаты представлены в таблице III.

Таблица III
Сравнение метрик. Для сравнения качества использовался AUC[3]

Метрика	AUC
WER	0.69
Расстояние Левенштейна	0.63
Лемматизированный WER (LER)	0.67
MERaLM	0.77

Можем заметить, что MERaLM действительно показала лучшее качество по сравнению с предыдущими метриками (WER, Расстояние Левенштейна) по основным характеристикам. Метрика LER, по сути, является частным случаем метрики MERaLM с единственным признаком — словарной формой слова. Тот факт, что LER показал результаты хуже, чем MERaLM говорит о том, что одного признака недостаточно. Но комбинация с другими, пусть даже незначительными признаками дает значительный прирост в качестве.

[1]https://github.com/gordeeva-ln/MERa/Коэффициенты.pdf

Таким образом, мы показали, что с помощью методов машинного обучения и правильной подготовки признаков можно отлавливать потери смысла. Разумеется данная конкретная модель не является оптимальной. Большее число признаков и более качественные методы обучения позволяют получать лучшие результаты. Но, наборы признаков будут зависеть от конкретной задачи и данных и наша цель была показать работоспособность подхода, с чем такой простой пример успешно справился.

VIII. Направление для дальнейших исследований

В данной работе мы предложили подход к построению семейства метрик. В качестве примера мы рассмотрели простую линейную модель, обобщающую метрику WER. Несмотря на тот факт, что в качестве модели выбран наиболее простой вариант, набор признаков выбран не оптимально, а на основе эмпирических предположений, нам удалось показать, что даже на таком наборе параметров существующие метрики оказываются хуже.

- В анализе естественного языка наилучшее качество в обширном спектре задач показывают нейросетевые модели на основе BERT [12] и XLNet [17]. Модели на основе нейронных сетей сложно интерпретировать, но для ситуации, когда требуется только предсказывать вероятность того, что распознавание передает смысл предложения, такие модели должны показывать очень хороший результат.
- Достоинство предложенной метрики MERa— хорошая интерпретируемость. Однако в рамках данной статьи она была использована в качестве демонстрации работы нового подхода и во многом нуждается в улучшениях. Для последующего использования, необходимо пересмотреть набор признаков и выбранную моделей для оценки параметров.
- Сбор данных для обучения метрики является одной из ключевых задач. От того, насколько полно данные описывают ситуации, которые важно распознать как одинаковые или разные по смыслу, зависит качество работы метрики. В зависимости от той предметной области, в которой метрика будет использована, этот набор данных окажется своим и всегда можно будет дообучиться.

IX. Заключение

В рамках этой статьи мы по-новому взглянули на метрики для задачи автоматического распознавания голоса и обнаружили, что существующие варианты не удовлетворяют базовым понятиям релевантности для данной задачи. Мы предложили новый подход, который, несмотря на свою простоту, позволяет отобразить насколько передает метрика информацию об эквивалентности гипотезы и исходного текста по смыслу. В качестве демонстрации применения этого подхода было предложено новое семейство метрик, которое строится на основе оценки передачи смысла. Проведенный эксперимент показал, что метрика из этого

семейства, реализованная на основе линейной модели оказывается лучше старых согласно нашему методу оценки качества метрик.

Список литературы

[1] В. И. Левенштейн(1965): *Двоичные коды с исправлением выпадений, вставок и замещений символов*, 4: 845-848.

[2] Dempster, A. P. / Laird, N. M. / Rubin, D. B.(1977): *Maximum likelihood from incomplete data via the EM algorithm*1-38.

[3] Bradley, Andrew P.(1997): *The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms*, 7: 1145–1159.

[4] Papineni, Kishore / Roukos, Salim / Ward, Todd / Zhu, Wei Jing(2002): *Bleu: a Method for Automatic Evaluation of Machine Translation*In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics311–318.

[5] Järvelin, Kalervo / Kekäläinen, Jaana(2002): *Cumulated Gain-Based Evaluation of IR Techniques*, 4: 422–446.

[6] Lin, Chin Yew / Och, Franz Josef(2004): *Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics*In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics605–es.

[7] Morris, Andrew Cameron / Maier, Viktoria / Green, Phil(2004): *From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition*2765-2768.

[8] Chan, William / Jaitly, Navdeep / Le, Quoc V. / Vinyals, Oriol(2015): *Listen, Attend and Spell*.

[9] Ali, A. / Magdy, W. / Bell, P. / Renais, S.(2015): *Multi-reference WER for evaluating ASR for languages with no orthographic rules*In: 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)576-580.

[10] Le, Ngoc Tien / Servan, Christophe / Lecouteux, Benjamin / Besacier, Laurent(2016): *Better Evaluation of ASR in Speech Translation Context Using Word Embeddings*In: Interspeech 2016.

[11] Pratap, Vineel / Hannun, Awni / Xu, Qiantong / Cai, Jeff / Kahn, Jacob / Synnaeve, Gabriel / Liptchinsky, Vitaliy / Collobert, Ronan(2018): *wav2letter++: The Fastest Open-source Speech Recognition System*.

[12] Devlin, Jacob / Chang, Ming Wei / Lee, Kenton / Toutanova, Kristina(2019): *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)4171–4186.

[13] Synnaeve, Gabriel u.a.(2019): *End-to-end ASR: from Supervised to Semi-Supervised Learning with Modern Architectures*.

[14] Li, Jason / Lavrukhin, Vitaly / Ginsburg, Boris / Leary, Ryan / Kuchaiev, Oleksii / Cohen, Jonathan M. / Nguyen, Huyen / Gadde, Ravi Teja(2019): *Jasper: An End-to-End Convolutional Neural Acoustic Model*.

[15] Lundberg, Scott M / Lee, Su In (2017): *A Unified Approach to Interpreting Model Predictions*. In: Guyon, I. / Luxburg, U. V. / Bengio, S. / Wallach, H. / Fergus, R. / Vishwanathan, S. / Garnett, R. (Hg.), Advances in Neural Information Processing Systems 30.Curran Associates, Inc.: 4765–4774.

[16] Mohamed, Abdelrahman / Okhonko, Dmytro / Zettlemoyer, Luke (2019): *Transformers with convolutional context for ASR*
.

[17] Yang, Zhilin / Dai, Zihang / Yang, Yiming / Carbonell, Jaime / Salakhutdinov, Russ R / Le, Quoc V (2019): *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. In: Wallach, H. / Larochelle, H. / Beygelzimer, A. / Alché Buc, F.d'/ Fox, E. / Garnett, R. (Hg.), Advances in Neural Information Processing Systems 32.Curran Associates, Inc.: 5754–5764.

Meaning Error Rate

Liudmila Gordeeva
ITMO University
lulu.gordeeva07@gmail.com

Vasily Ershov
Yandex
noxoomo@yandex-team.ru

Igor Labutin
Yandex / SPb HSE
Labutin.IgorL@gmail.com

Igor Kuralenok
Yandex / JetBrains Research
solar@yandex-team.ru

Abstract—Currently, WER (Word Error Rate) is used as a metric for automatic speech recognition systems quality evaluation. This metric is rather simple and works well in many cases. But, WER and similar metrics do not take into account several key factors. The most critical one is a distortion of a phrase meaning by speech recognition systems, WER metric is not sensitive for such error types. Besides, some applications do not need perfect recognition word to word. Their specific requirements may be: identifying current user intent by voice assistants; exact recognition of licence plate, address, phone number, etc. To estimate the quality of speech recognition systems satisfying such requirements a new metric should be designed. One, that will reflect not only errors in words but also a semantic distortion.

Here we present a new general approach for the construction of a metric. The main idea of this approach is using crowdsourcing on the first stage of collecting a dataset and the next reduction of the construction problem to a well known supervised learning task. As an application example, we propose the generalization of the WER — MERaLM metric, which has the following advantages: it considers that different mistakes affect the meaning distortion in different ways, and; the easy interpretability of the assessment.

Index Terms—automatic speech recognition, machine learning