УДК 004.934:811.512.122

Болатов С.
учащийся 10 класса
школа «Физтех»
Казахстан, г. Алматы
Турсын Г.
ментор
образовательная программа ADMIT
Казахстан

СКОЛЬКО ЧАСОВ КАЗАХСКОЙ РЕЧИ НУЖНО МОДЕЛИ РАСПОЗНАВАНИЯ WHISPER: ОЦЕНКА ПО 59 ЯЗЫКАМ ДЛЯ МОДЕЛЕЙ РАЗНОГО РАЗМЕРА

Аннотация: Модель распознавания речи Whisper обучалась всего на 12 часах казахской речи и ошибается в казахском намного чаще, чем в русском. По опубликованным данным о числе часов обучения и доле ошибочных слов (WER) для 45–59 языков построена зависимость ошибки от объёма данных для каждого из шести размеров модели и оценено, сколько часов казахской речи понадобилось бы для заданного уровня ошибок. При объёме открытого казахского корпуса ISSAI KSC2 (около 1 130 часов) модель small (244 млн параметров), которую можно запускать на обычном компьютере, по этой зависимости ошибалась бы примерно в 21% слов вместо 70% (28%, если сохранится отставание казахского от общей зависимости), крупнейшая модель – примерно в 10% вместо 38%. Маленькие модели получают от дополнительных данных меньше: каждое удвоение часов снижает ошибку на 10% у самой маленькой модели и на 16% у крупнейшей.

Ключевые слова: распознавание речи, казахский язык, Whisper, малоресурсные языки, объём обучающих данных, языковые модели

Bolatov S.
10th-grade student
Fizmat school "Fiztekh"
Kazakhstan, Almaty
Tursyn G.
mentor
ADMIT educational program
Kazakhstan

HOW MANY HOURS OF KAZAKH SPEECH DOES THE WHISPER RECOGNITION MODEL NEED: AN ESTIMATE FROM 59 LANGUAGES FOR MODELS OF DIFFERENT SIZES

Abstract: The Whisper speech recognition model was trained on only 12 hours of Kazakh speech and makes far more errors in Kazakh than in Russian. Using published data on training hours and word error rate (WER) for 45–59 languages, the dependence of error on data volume was estimated for each of six model sizes, and the number of hours of Kazakh speech needed for a given error level was derived. At the size of the open Kazakh corpus ISSAI KSC2 (about 1,130 hours), the small model (244 million parameters), which can run on an ordinary computer, would by this dependence misrecognise about 21% of words instead of 70% (27% if Kazakh's gap to the common trend persists), and the largest model about 10% instead of 38%. Small models gain less from additional data: each doubling of hours lowers the error by 10% for the smallest model and by 16% for the largest.

Keywords: speech recognition, Kazakh language, Whisper, low-resource languages, training data volume, language models

Введение

Впервые автор заметил, что нейросеть плохо понимает казахский язык, когда помогал младшему брату с домашним заданием: перевод с казахского получился странным и бессмысленным, а с русским и английским такого не происходило. В предыдущей работе автора, направленной в Central Asian Scientific Journal, по опубликованным данным модели распознавания речи Whisper [1] показано, что казахский язык переходит границу в 50% ошибочных слов только на модели среднего размера и что его отставание связано прежде всего с малым объёмом обучающих данных: в обучении было 12 часов казахской речи против 9 761 часа русской.

Если дело в объёме данных, возникает практический вопрос: сколько данных нужно. Открытый казахский речевой корпус ISSAI KSC2 содержит около 1 128 часов расшифрованной речи [2], то есть почти в сто раз больше, чем было у Whisper. Настоящая статья оценивает, какой уровень ошибок можно было бы ожидать при таком объёме данных, и отдельно рассматривает маленькие модели, которые можно запускать на обычном компьютере без интернета.

Методы

Использованы опубликованные авторами Whisper данные: доля ошибочных слов (WER, %) на тестовом наборе FLEURS [3] для 82 языков и шести размеров модели – tiny, base, small, medium, large и large-v2 (табл. 13 статьи [1]) – и число часов расшифрованной речи каждого языка в обучающих данных (рис. 11 статьи [1]). Числа извлечены из PDF статьи программно и проверены по опубликованному в статье коэффициенту детерминации (0,83).

Для каждого размера модели по всем языкам, кроме казахского, с известным числом часов и WER ниже 100% построена линейная зависимость логарифма WER от логарифма часов обучения (от 45 языков для tiny до 59 для large-v2). По этой зависимости рассчитаны ожидаемая WER казахского при 1 128 часах и число часов, при котором ожидаемая WER опускается ниже 30, 20 и 15%. Неопределённость оценена бутстрепом по языкам (2 000 повторений). Так как казахский язык в Whisper ошибается чаще, чем предсказывает зависимость для его 12 часов (в 1,24–1,69 раза в зависимости от размера модели), приведён и второй вариант расчёта, в котором это отставание сохраняется.

Результаты

Зависимость ошибки от часов обучения хорошо описывает все размеры моделей, кроме самого маленького: коэффициент детерминации от 0,72 до 0,80 для base–large-v2 и 0,64 для tiny. При 12 часах эта зависимость предсказывает для казахского WER 53,1% у модели small и 30,3% у large-v2; фактические значения выше – 70,3% и 37,7%.

При объёме корпуса KSC2, около 1 128 часов, ожидаемая WER казахского по общей зависимости составляет 20,9% для модели small (95% ДИ 18,6–23,5%), 13,0% для medium и 9,9% для large-v2 (табл. 1). Если отставание казахского от общей зависимости сохранится, ожидаемые значения выше: 27,7%, 16,1% и 12,3% соответственно. Для самой маленькой модели tiny даже 1 128 часов дают около 50%, а с учётом отставания – 84%.

Таблица 1. Казахский язык в Whisper: фактическая ошибка и оценка по зависимости от часов обучения

Модель (млн параметров)	**WER казахского сейчас (12 ч), %**	**Ожидаемая WER при 1 128 ч, % (95% ДИ)**	**То же с поправкой на отставание казахского,**	**Часов для WER ниже 20%**	**Снижение WER при удвоении часов, %**

			%		
tiny (39)	165,2	49,9 (45,6–54,7)	84,3	вне данных	9,8
base (74)	109,2	35,9 (32,2–39,9)	47,6	≈ 28 000	11,9
small (244)	70,3	20,9 (18,6–23,5)	27,7	≈ 1 400 (810–2 760)	13,3
medium (769)	48,8	13,0 (11,6–14,7)	16,1	≈ 190 (140–290)	15,5
large (1 550)	43,8	11,7 (10,1–13,7)	14,9	≈ 120 (80–180)	15,1
large-v2 (1 550)	37,7	9,9 (8,6–11,5)	12,3	≈ 65 (43–99)	15,7

Чтобы модель small ошибалась меньше чем в 20% слов, по общей зависимости нужно около 1 400 часов казахской речи (95% ДИ 810–2 760), а с учётом отставания казахского – около 5 500 часов. Для ошибки ниже 30% достаточно около 190 часов (около 760 с учётом отставания). Для крупнейшей модели large-v2 граница в 20% достигается уже примерно при 65 часах.

Маленькие модели получают от данных меньше. Каждое удвоение часов обучения снижает WER на 9,8% у модели tiny, на 13,3% у small и на 15,5–15,7% у medium и large-v2. Поэтому разрыв между маленькими и большими моделями с ростом данных не сокращается, а в относительном выражении растёт.

Обсуждение

Автор объясняет это так. У маленькой модели мало параметров, поэтому она может вместить меньше знаний, чем большая. Когда маленькая модель получает новые данные, она усваивает из них меньше, потому что её место почти заполнено. Большая модель может извлечь из

тех же данных больше пользы, поэтому удвоение часов снижает её ошибку на 16%, а у самой маленькой модели только на 10%.

Из результатов следуют три вывода. Во-первых, данные, которые уже существуют в Казахстане, по порядку величины соответствуют тому, что нужно: при объёме корпуса KSC2 ожидаемая ошибка модели среднего размера – около 13–16%, что близко к уровню, при котором распознаванием можно пользоваться. Во-вторых, для маленьких моделей, которые можно запускать без интернета, одного объёма данных, вероятно, недостаточно: даже 1 128 часов дают модели small около 21–28% ошибок, и каждое удвоение данных помогает ей меньше, чем большим моделям. В-третьих, казахский язык в Whisper ошибается чаще, чем предсказывают его 12 часов, поэтому реальная потребность в данных может быть ближе к верхним оценкам.

Ограничения работы. Зависимость построена по разным языкам, а не по казахскому: она показывает, как ошибка связана с объёмом данных у языков в целом, но число часов связано и с письменностью, и с тем, сколько текста на языке есть в интернете, поэтому это связь, а не доказанная причина. Расчёт описывает модель, обученную с нуля вместе со всеми языками, а не дообучение готовой модели на казахских данных, которое на практике может дать другой результат. Часы в статье Whisper указаны округлённо, а для 15 языков, включая английский, не указаны. Новые версии модели (large-v3) в опубликованной таблице отсутствуют.

Заключение

По опубликованным данным Whisper, около 1 130 часов казахской речи, то есть объём уже существующего открытого корпуса ISSAI KSC2, по общей зависимости снизили бы ошибку модели среднего размера примерно с 49% до 13–16%, а маленькой модели, которую можно запускать на обычном компьютере, – с 70% до 21–28%. Для маленьких моделей одного увеличения данных, вероятно, недостаточно, так как

каждое удвоение данных помогает им меньше, чем большим. Включение открытых казахских корпусов в обучение многоязычных моделей – наиболее прямой путь к тому, чтобы казахский язык распознавался не хуже русского.

Вклад авторов и использование ИИ. С. Болатов написал первый абзац введения о случае с младшим братом и объяснение в разделе «Обсуждение» о том, почему маленьким моделям данные помогают меньше; предыдущая работа С. Болатова, на данных которой основана статья, исходила из его прогноза. Он самостоятельно сверил с таблицей 13 и рисунком 11 работы [1] ошибку Whisper в казахском языке и объём казахских обучающих данных, при сверке с работой [2] обнаружил, что объём корпуса KSC2 составляет около 1 128 часов, а не 1 200, как было указано в первой версии статьи, и вручную пересчитал по полученной зависимости ожидаемую ошибку модели small. Г. Турсын (ментор образовательной программы ADMIT) сформулировал вопрос статьи, выполнил расчёты, написал программный код и текст с помощью ИИ-ассистента (Claude). Оба автора прочитали статью и согласны с ней.

Использованные источники:

1. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision // Proceedings of the 40th International Conference on Machine Learning. – PMLR, 2023. – Vol. 202. – P. 28492–28518. URL: https://arxiv.org/abs/2212.04356 (дата обращения: 26.10.2026).

2. Mussakhojayeva S., Khassanov Y., Varol H.A. KSC2: An industrial-scale open-source Kazakh speech corpus // Proceedings of Interspeech 2022. – 2022. – P. 1367–1371. URL: https://www.isca-archive.org/interspeech_2022/mussakhojayeva22_interspeech.html (дата обращения: 26.10.2026).

3. Conneau A., Ma M., Khanuja S., Zhang Y., Axelrod V., Dalmia S., Riesa J., Rivera C., Bapna A. FLEURS: Few-shot learning evaluation of universal representations of speech // 2022 IEEE Spoken Language Technology Workshop (SLT). – 2023. – P. 798–805. DOI: 10.1109/SLT54892.2023.10023141.