УДК 004

Шамало В.В.

магистрант кафедры информатики и вычислительной техники

Кубанский государственный технологический университет

Россия, г. Краснодар

КАК СОБРАТЬ РАЗМЕЧЕННЫЙ КОРПУС КОРОТКИХ РУССКОЯЗЫЧНЫХ ТЕКСТОВ: МЕТОДИКА И ТИПИЧНЫЕ ОШИБКИ

Аннотация: существующие датасеты коротких текстов почти не подходят для русского языка: они собраны под другую задачу или получены машинным переводом, что портит качество. В данной статье описано, как собирали корпус NRS-H из 116 тысяч реплик, где соблюден баланс классов. Также разобраны четыре шага — отбор источников, очистка от ботов, балансировка классов и проверка разметки, а также ошибки, которые всплыли уже после обучения моделей. Основная проблема заключалась в очистке данных: спам отфильтровали, но типовые вежливые ответы модераторов оставили, из-за чего модель интерпретировала их как комментарии, сгенерированные моделью.

Ключевые слова: корпус текстов, разметка данных, короткие тексты, русский язык, боты, балансировка классов.

Shamalo V.V.

Master's Student

Kuban State Technological University

Russia, Krasnodar

HOW TO BUILD AN ANNOTATED CORPUS OF SHORT RUSSIAN TEXTS: A METHOD AND ITS TYPICAL ERRORS

Abstract: open datasets of short texts are of little use for Russian: they were collected for other tasks or obtained by machine translation. The paper describes how the NRS-H corpus of 116,000 utterances, evenly split between human and

machine ones, was built. Four steps are covered — choosing sources, cleaning out bots, balancing classes and checking the labels — together with the errors that surfaced only after the models had been trained. The main one: spam was removed while routine polite replies from moderators were kept, and the model took them for machine-generated.
Keywords: text corpus, data labelling, short texts, Russian language, bots, class balancing.

Введение

Чтобы обучить классификатор коротких реплик, нужен корпус данных, где эти реплики собраны в естественном виде. Открытые датасеты такой возможности почти не дают. Одни датасеты собраны под другую задачу и содержат длинные тексты, другие переведены машиной, и в переводе исчезает именно то, что отличает живую речь от машинной, а именно сбитый порядок слов, сокращения и опечатки.

Отдельная сложность в том, что сама реплика длиной пять–пятнадцать слов почти не несёт признаков сама по себе. В длинном тексте мы можем видеть построение абзаца, переходы, повторы, а вот в короткой реплике всего этого нет, и любая примесь в данных сразу сдвигает модель в область неточности.

Свой корпус собрать можно, но на этом пути легко испортить данные так, и заметно это станет только после обучения. В этой статье описано, как собирали корпус NRS-H, и какие ошибки при этом были допущены. Опыт пригодится всем, кто размечает короткие тексты на русском.

Методы и исследования

Данный процесс разбили на четыре шага: выбрать источники, убрать ботов, привести к балансу классов и проверить разметку. Ниже каждый шаг описан подробно вместе с решениями, которые пришлось принимать по ходу.

Результаты

Шаг первый: откуда брать реплики. Человеческую половину корпуса набирали из комментариев в открытых сообществах ВКонтакте. Важно было взять разные темы: новости, киберспорт, наука, развлечения. Если ограничиться одной темой, модель выучит лексику этой темы, а не признаки живой речи. А вот машинную половину сгенерировали пятью языковыми моделями — GPT-4, Claude 3, Gemini 2.5, YandexGPT 3 и GigaChat. Важно понимать, что одной модели мало, так как детектор быстро привыкает к её манере и если дать текст другой модели, то он не справится. Проверить это просто, можно обучить на текстах одной модели и протестировать на текстах другой и падение точности видно сразу.

Сколько нужно данных. Итоговый объём составил 116 тысяч реплик. На выборках меньше пятидесяти тысяч точность заметно скакала от запуска к запуску, дальше стабилизировалась. Наращивать корпус бесконечно смысла нет — прирост быстро упирается в качество разметки, а не в объём. И разумнее потратить силы на чистоту данных, чем на удвоение их количества.

Шаг второй: очистка от ботов. Комментарии ботов попадают в корпус под видом человеческих и засоряют его. Их отсеивали по нескольким признакам: аккаунт дублирует сообщения в разных сообществах, пишет с аномальной частотой, а создан совсем недавно и почти не имеет истории. Признаки известны и описаны в литературе [2], но ни один из них по отдельности не работает: у живого человека тоже бывает новый аккаунт, а у активного комментатора может быть сотня сообщений в день. Сработало сочетание этих признаков, а окончательное решение по спорным аккаунтам принимали уже вручную. Заодно убрали реплики, состоящие из одних эмодзи, и подписи к медиавложениям: смысла в них нет, а классу «человек» они добавляют шум.

Шаг третий: поровну обоих классов. Классы уравняли по числу примеров, 50 на 50. Если оставить перекос, модель научится угадывать преобладающий класс и покажет неплохую точность, не научившись ничему полезному. Уравнивать можно и синтетически, дописывая примеры редкого класса [3], но с короткими текстами метод не будет работать, алгоритм просто не выдаст осмысленную реплику из пяти слов. Проще собрать больше живых данных.

Шаг четвёртый: проверка разметки. Часть корпуса подверглась двойной независимой аннотации и оценили уровень согласия разметчиков. Данный параметр нужен, так как без него невозможно определить, чем ограничен предел точности или моделью, или всё-таки исходными данными [4]. Наибольшие расхождения выявлены в ожидаемых сегментах: короткие клише вроде «спасибо, учтём» часть аннотаторов классифицировала как человеческие, а часть — как машинные.

Что пошло не так. Первая ошибка выявилась на этапе тестирования. Из человеческого сегмента данных исключили спам, но сохранили стандартизированные ответы модераторов. Модель ошибочно классифицировала их как машинные, что привело к росту ложноположительных срабатываний. Вторая проблема возникла из-за корректировки данных без пересчёта баланса классов после удаления сообщений, состоящих только из эмодзи. Третья ошибка заключалась в использовании однотипных вводных конструкций при генерации текстов: модель заучила конкретную формулировку вместо общего стиля. Подобные артефакты («ярлыки») в данных выявляются преимущественно при ручном анализе ошибок [5].

Заключение

В работе описана методика сбора корпуса коротких русскоязычных реплик: разноплановые источники человеческой речи, несколько языковых моделей для машинной, отсеивание ботов по комплексу признаков, равный

баланс классов и двойная разметка контрольной выборки. Самыми затратными оказались ошибки не сбора, а чистки: шаблонные фразы людей модель приняла за автоответы, так как их не проверили отдельно. Главный практический вывод — разбирать вручную нужно не случайную выборку, а именно сбои модели. В дальнейшем корпус стоит пополнить текстами свежих поколений LLM и проверить устойчивость детектора к перефразированию [6].

Использованные источники:

1. Gebru T., Morgenstern J., Vecchione B. [et al.] Datasheets for Datasets // Communications of the ACM. 2021. Vol. 64. № 12. P. 86–92.

2. Ferrara E., Varol O., Davis C. [et al.] The Rise of Social Bots // Communications of the ACM. 2016. Vol. 59. № 7. P. 96–104.

3. Chawla N.V., Bowyer K.W., Hall L.O. [et al.] SMOTE: Synthetic Minority Over-sampling Technique // Journal of Artificial Intelligence Research. 2002. Vol. 16. P. 321–357.

4. Artstein R., Poesio M. Inter-Coder Agreement for Computational Linguistics // Computational Linguistics. 2008. Vol. 34. № 4. P. 555–596.

5. Northcutt C.G., Athalye A., Mueller J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks // Proceedings of the NeurIPS Datasets and Benchmarks Track. 2021.

6. Kuratov Y., Arkhipov M. Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language // Computational Linguistics and Intellectual Technologies. 2019. P. 333–340.