https://doi.org/10.17323/jle.2024.22272

Борьба с инфляцией оценок: концентрированные датасеты в задаче автоматического исправления ошибок

Старченко Владимир [1], Харламова Дарья [1], Клыкова Елизавета [2], Шаврина Анастасия [1], Старченко Алексей [1], Виноградова Ольга [2], Ляшевская Ольга [1,3]

[1]Национальный исследовательский университет Высшая школа экономики, г. Москва, Россия

[2]Независимый исследователь

[3]Институт русского языка им. В. В. Виноградова РАН, г. Москва, Россия

АННОТАЦИЯ

Введение: Системы автоматического исправления ошибок (Grammatical Error Correction, GEC) значительно продвинулись за последнее десятилетие. Согласно стандартным метрикам, они часто достигают уровня экспертов-людей или даже превосходят их. Тем не менее, эти системы плохо справляются с некоторыми типами ошибок, которые легко исправляются носителями языка. Таким образом, алгоритмы оценки и оценочные наборы данных достигли предела своей разрешающей способности и больше не позволяют совершенствовать GEC-системы

Цель: Решить проблему ограничения разрешающей способности в задаче GEC. Предлагаемый подход заключается в использовании для оценки концентрированных наборов данных с повышенной плотностью ошибок, представляющих сложность для современных GEC-систем.

Методология: Для проверки предложенного решения рассматривались ошибки, требующие учета широкого контекста. Был разработан концентрированный датасет для английского языка с высокой плотностью таких ошибок путем полуавтоматического объединения размеченных примеров из четырех источников и расширения аннотации ошибок, связанных с широким контекстом. На основе полученного датасета были провалидированы две GEC-системы. Были использованы как традиционные алгоритмы оценки, так и новый подход, модифицированный для учета более широкого контекста.

Результаты: Концентрированный датасет включает 1014 примеров, отобранных вручную из датасетов FCE, CoNLL-2014, BEA-2019 и REALEC. Он содержит разметку нескольких типов контекстно зависимых ошибок, таких как ошибки в местоимениях, глагольном времени, пунктуации, референциальных средствах и коннекторах. GEC-системы демонстрируют более низкие результаты при оценке на датасете с высокой плотностью выбранных типов ошибок по сравнению со случайным набором данных с аналогичными параметрами.

Заключение: Снижение метрик при оценке на концентрированных наборах данных подтверждает ценность таких датасетов для совершенствования GEC-систем. Разработанный датасет может лечь в основу дальнейших исследований в области автоматического исправления ошибок, требующих учета широкого контекста.

КЛЮЧЕВЫЕ СЛОВА:

Исправление грамматических ошибок, ошибки изучаемого иностранного языка (L2), английский как второй язык (ESL), концентрированные наборы данных, система исправления грамматических ошибок кросс-предложений

Для цитирования: Старченко, В., Харламова, Д., Клыкова, Е., Шаврина, А., Старченко, А., Виноградова, О., & Ляшевская, О. (2024). Борьба с инфляцией оценок: концентрированные наборы данных для исправления грамматических ошибок. *Journal of Language and Education, 10*(4), 115-133. https://doi.org/10.17323/jle.2024.22272

Correspondence:
Старченко Владимир,
vmstarchenko@edu.hse.ru

Получена: 18 августа 2024

Принята: 16 декабря 2024

Опубликована: 30 декабря 2024

ВВЕДЕНИЕ

Автоматическое исправление ошибок (Grammatical Error Correction, GEC) является важной задачей обработки естественного языка (Natural Language Processing, NLP). Оно включает выявление и исправление орфографических и пунктуационных ошибок, изменение синтаксических структур, а также подбор подходящей лексики и определение порядка слов для улучшения читаемости и ясности текста. Эта задача охватывает не только обнаружение, классификацию и исправление форм и структур, которые являются «строго грамматическими

по своей природе» (Bryant et al., 2023), но и более глубокий контекстный анализ и улучшение качества текста, что обеспечивает соответствие исправлений замыслу и стилю текста (Du & Hashimoto, 2023). Использование технологии GEC может быть полезным в обучении, как для детей, так и для изучающих иностранный язык (L2). Кроме того, эти технологии экономят время преподавателей языка, а также оптимизируют работу корректоров, редакторов и других специалистов, работающих с текстами.

За последнее десятилетие GEC-системы достигли значительных успехов. Qorib and Ng (2022) отмечают, что современные модели, такие как GECToR (Omelianchuk et al., 2020) и T5 (Rothe et al., 2021), по стандартным метрикам даже превосходят экспертов-людей. Однако эти системы по-прежнему не способны выявлять и исправлять некоторые ошибки, которые не представляют сложности для образованного носителя языка. Таким образом, задача GEC столкнулась с кризисом предела разрешающей способности метрик: несмотря на то, что потенциал для улучшения обработки отдельных типов ошибок сохраняется, сами метрики достигли своего предела.

До настоящего времени решения данной проблемы предложено не было. Одним из возможных подходов является раздельная оценка моделей по различным типам ошибок. Это позволяет GEC-системам выявлять более сложные типы ошибок, но не решает проблему недостаточной представленности сложных ошибок в существующих датасетах. Кроме того, такая практика редко встречается в исследованиях (см. Yuan & Bryant, 2021; Zhang et al., 2022) и, что важнее, не используется для дальнейшего сравнения и обучения моделей. Данное исследование делает шаг к решению проблемы предела разрешающей способности для оценки качества автоматического исправления ошибок.

Одним из типов ошибок, на которой влияет эта проблема, являются ошибки, для обнаружения или исправления которых требуется информация из широкого контекста (т. е. контекста, выходящего за пределы одного предложения). В литературе существует консенсус относительного того, что такие ошибки особенно сложны для моделей как по техническим причинам (например, из-за распространенной практики обучения моделей на уровне предложений, а не текстов), так и из-за сложности учета дальних зависимостей (Chollampatt et al., 2019; Yuan & Bryant, 2021; Qorib & Ng, 2022). Предел разрешения затрудняет прогресс GEC-систем в отношении контекстно зависимых ошибок при использовании стандартных оценочных датасетов и метрик.

В настоящем исследовании предлагается использовать оценочные датасеты с более высокой долей ошибок, представляющих сложность для GEC-систем. Ожидается, что такие датасеты позволят получить более низкие значения метрик и обеспечат возможность дообучения моделей для исправления сложных ошибок. В качестве материала для проверки этой гипотезы были выбраны ошибки, связанные с широким контекстом. Разработанный концентрированный набор данных содержит 1014 примеров, отобранных из широко используемых датасетов для задачи GEC-. Он состоит из дополнительно размеченных примеров, каждый из которых содержит как минимум одну ошибку, требующую широкого контекста для ее исправления. Чтобы проверить, обеспечивает ли концентрированный набор данных более высокую разрешающую способность, мы оценили работу двух нейронных сетей (BART и T5) на созданном наборе данных. Было показано, что обе модели демонстрируют низкие метрики при оценке на концентрированном датасете, несмотря на то, что для задачи GEC в целом результаты являются достаточно высокими. Таким образом, созданный датасет прокладывает путь для улучшения качества GEC-систем, так как более низкие (но более точные) результаты метрик делают оценивание более дифференцированным и открывает возможности для улучшения моделей. Данное исследование также вносит вклад в область применения методов машинного обучения к решению проблемы дальних зависимостей, предоставляя инструмент для оценки и сравнения моделей с учетом контекстно зависимых ошибок.

ОБЗОР ЛИТЕРАТУРЫ

Задача автоматического исправления ошибок

Попытки автоматизированного исправления ошибок предпринимаются с самых ранних этапов развития компьютерных технологий. Первые практически ориентированные исследования были сосредоточены на исправлении орфографических ошибок (Cargill, 1980; Bentley, 1985), в то время как задача GEC в более широком смысле обсуждалась в основном как этап предобработки для NLP-систем, неспособных обрабатывать грамматически некорректные входные данные (Kwasny & Sondheimer, 1981; Jensen et al., 1983). Первые инструменты GEC, созданные для практического применения, появились позже (Burstein et al., 2003; Leacock et al., 2009 и др.), большинство из них опирались на подходы, основанные на правилах.

Практически ориентированные системы быстро перешли к методам машинного обучения с учителем. Эти методы основаны на классификации (Lee, 2004; Rozovskaya & Roth, 2010; Dahlmeier & Ng, 2011) и архитектурах статистического машинного перевода (Brockett et al., 2006; Yuan & Felice, 2013). Подробный обзор исследований, посвященных GEC на этом этапе, представлен в работе Leacock et al. (2014).

С появлением методов глубокого обучения и больших языковых моделей (LLM) в развитии GEC-систем произошел рывок. Были разработаны системы на основе различных архитектур, включая рекуррентные нейронные сети (RNN

— см. Yuan & Briscoe, 2016; Xie et al., 2016; Wang et al., 2017), свёрточные нейронные сети (CNN — см. Chollampatt & Ng, 2018) и трансформеры (Edunov et al., 2018; Wang et al., 2019 и др.). Более детальный анализ недавних достижений в области GEC представлен в работах Wang et al. (2021) и Bryant et al. (2023).

Предел разрешающей способности в оценке GEC-систем

Современные GEC-системы, скорее всего, достигли предела разрешающей способности: по распространенным метрикам (таким, как $F_{0,5}$, предложенной для задачи GEC в работе Ng et al. (2014)) они неоднократно превосходили результаты экспертов-людей (Qorib & Ng, 2022). Однако в более поздних исследованиях (Zhou et al., 2023; Li & Wang, 2024) говорится о возможности дальнейших улучшений для выполнения задачи автоматического исправления ошибок.

Здесь необходимо сделать два важных замечания. Во-первых, говоря о «низких» оценках аннотаторов, мы не имеем в виду, что они вызваны неточностями экспертной разметки. Неполное согласие аннотаторов в основном связано с тем, что эксперты выбирают разные, но одинаково допустимые варианты исправления ошибок в исходном тексте.

Во-вторых, нельзя утверждать, что результаты автоматического исправления ошибок идеальны. Qorib and Ng (2022, стр. 2795–2797) перечисляют несколько типов ошибок, которые GEC-модели систематически пропускают. Среди них такие распространенные классы ошибок, как неточности в синтаксических структурах (например, в согласовании подлежащего и сказуемого); ошибки в длинных предложениях; предложения с большим количеством ошибок; ошибки, затрагивающие более одного предложения; ошибки, требующие перефразирования части предложения (например, исправление фраз, которые звучат неестественно) и др. Еще один яркий пример типа ошибок, который не вызывает трудностей у носителей языка, но регулярно обсуждается как проблемный для современных систем GEC, это орфографические ошибки (Chollampatt & Ng, 2018; Starchenko & Starchenko, 2023).

Вероятное объяснение этих недостатков GEC-систем заключается в том, что количество сложных ошибок в обучающих и тестовых датасетах недостаточно велико, чтобы серьезно повлиять на метрики. Тем не менее такие ошибки достаточно заметны и при ручной разметке, и при применении GEC-систем (например, обсуждение ошибок на уровне символов в работе Starchenko & Starchenko (2023)). Это может быть связано с тем, что обучающие и тестовые датасеты создаются на основе текстов, написанных людьми, которые не являются носителями языка, и часто перегружены базовыми грамматическими ошибками.

Некоторые исследователи публикуют результаты оценки для различных типов ошибок по отдельности (Yuan & Bryant, 2021; Zhang et al., 2022), что дает возможность отдельно оценивать эффективность модели в отношении ошибок, исправляемых хуже других. Эта практика становится все более распространенной, особенно после появления инструмента ERRANT (Bryant et al., 2017), в котором реализован такой подход к оценке. Однако в настоящее время нет алгоритмов, которые напрямую используют такие раздельные оценки для обучения моделей. Обычно эти оценки обсуждаются как заключительные наблюдения, а не применяются для дообучения моделей.

Концентрированные датасеты в NLP

Концентрированные датасеты успешно применяются в различных областях NLP помимо задачи автоматического исправления ошибок. Один из примеров — обработка этических предвзятостей (biases) в больших языковых моделях. Несмотря на то что такие предвзятости достаточно редко встречаются в естественных данных, даже единичное их проявление в выходных данных оказывает значительное влияние на возможность коммерческого использования модели. В связи с этим модели дополнительно дообучаются и оцениваются на концентрированных датасетах, содержащих предвзятые данные (Nangia et al., 2020; Zhao et al., 2023).

В области GEC концентрированные датасеты пока не получили широкого распространения как инструменты оценки и обучения моделей. Starchenko (2024) создал синтетический концентрированный датасет для дообучения большой языковой модели для задачи GEC, а Starchenko and Starchenko (2023) предложили синтетический оценочный датасет. Однако оба исследования ограничились орфографическими ошибками, которые являются, вероятно, наиболее базовым типом вышеупомянутых сложных ошибок, что позволяет легко генерировать пары «ошибка — исправление». Chollampatt et al. (2019) создали синтетический датасет с ошибками в использовании глагольных времен. Насколько нам известно, до настоящего момента ни один концентрированный датасет, основанный на естественном языковом материале, не использовался в задаче GEC. Данное исследование заполняет этот пробел путем создания и применения концентрированного оценочного датасета, включающего размеченные примеры из нескольких корпусов текстов для изучающих язык как L2.

Зависимость от контекста в GEC

Проблема зависимости от контекста играет ключевую роль в задаче GEC. Обнаружение грамматической ошибки и способ ее исправления существенно зависят от контекста (например, от частей речи соседних слов, их лексической семантики и порядка). Тип дискурса и общий замысел автора также важны для правильного понимания требуемого исправления.

С самого начала GEC-системы в значительной степени полагались на контекст, что достигалось либо путем передачи некоторых его характеристик в модель (для классификаторов), либо за счет использования архитектур, учитывающих чувствительность к контексту (SMT, RNN, CNN, трансформеры). Тем не менее больше внимания обычно уделяется локальному контексту. Модели часто обучаются исправлять предложения вне широкого контекста (например, передовая модель Rothe et al. (2021, стр. 703–704)). Более того, многие широко используемые в задаче GEC датасеты содержат отдельные предложения, а не целые абзацы или тексты (например, относительно новый датасет JFLEG (Napoles et al., 2017)). В результате даже самые совершенные современные модели зачастую не способны исправить некоторые типы ошибок, чувствительные к широкому контексту (например, ошибки в местоимениях, времени и модальности глаголов и использовании дискурсивных маркеров).

Лишь немногие исследователи уделяют достаточное внимание более широкому контексту в GEC. Проблема контекстно зависимых ошибок обычно формулируется как «ошибки, которые требуют контекста из другого предложения для их исправления» (Qorib & Ng, 2022). Chollampatt et al. (2019) создали CNN-модель с дополнительным энкодером, сохраняющим информацию из предыдущих предложений, и внедрили это кодирование в декодер с использованием механизмов внимания. Yuan and Bryant (2021) сравнили различные архитектуры на основе трансформеров, оценив их производительность на ошибках, чувствительных к более длинному контексту.

МЕТОДОЛОГИЯ

Рабочее определение ошибок, чувствительных к широкому контексту

В данном исследовании рассматривается более узкая область практического применения, которая касается ошибок, исправление которых требует учета широкого контекста. Наиболее очевидный пример таких ошибок — это ошибки, требующие материала из другого предложения (т. н. межфразовые ошибки, cross-sentence errors). Рассмотрим дискурс, представленный в примере (1):

> (1) I go for a walk to a park every day with my two lovely Corgi dogs. I **met**→**[meet]** many people in the park.

Второе предложение в примере (1) грамматически корректно при изолированном рассмотрении, но из контекста первого предложения ясно, что глагол во втором предложении не может быть употреблен в прошедшем времени (Past Simple) и должен быть в настоящем времени (Present Simple). Как уже обсуждалось, такие примеры представляют трудности для современных нейросетевых моделей. Далее мы будем называть подобные ошибки контекстно зависимыми, подразумевая широкий, а не локальный контекст.

Заметим, что ошибки, чувствительные к контексту, могут также возникать в пределах одного предложения. Модификация (1') почти не отличается от примера (1), и контекст, необходимый для исправления ошибки во времени, практически идентичен:

> (1') I go for a walk to a park every day with my two lovely Corgi dogs, and I **met**→**[meet]** many people in the park.

Как будет показано далее, сложность таких ошибок для GEC-систем обусловлена не собственно границами предложений, поскольку случаи, подобные (1'), тоже представляют для них проблему. Поэтому важно учитывать ошибки, которые зависят от широкого контекста, расположенного внутри одного предложения. Это стирает четкую границу между локальным и широким контекстом в GEC: она не может быть проведена по границе между предложениями.

Для разметки примеров в датасете мы используем следующее рабочее определение:

Определение: Ошибки, которые невозможно обнаружить или исправить без учета материала из другой клаузы, содержащей личный (финитный) глагол, являются контекстно зависимыми или чувствительными к широкому контексту.

Таким образом, в датасет включаются не только ошибки, требующие доступа к другим предложениям, но и «межклаузальные» (cross-clause) ошибки. Очевидно, что это определение исключает некоторые возможные (и, вероятно, более спорные) случаи контекстно зависимого исправления ошибок, где широкий контекст находится в пределах одной клаузы или в другой нефинитной клаузе. Однако предложенное определение позволяет включить только бесспорные случаи контекстно зависимых ошибок, не ограничиваясь примерами, требующими для исправления материала из другого предложения. Более теоретически обоснованное определение ошибок, чувствительных к контексту, мы оставляем для будущих исследований.

Создание концентрированного датасета с повышенной долей контекстно зависимых ошибок

Концентрированный датасет с увеличенной частотой «межклаузальных» ошибок был создан на основе данных из существующих аннотированных датасетов. В этом разделе рассматривается алгоритм его создания и свойства четырех датасетов, из которых он составлен. Характеристики полученного датасета представлены в разделе «Результаты».

Характеристики используемых неконцентрированных датасетов

Концентрированный датасет включает примеры, аннотированные для задачи GEC, из следующих четырех источников:

1. **The First Certificate in English (FCE)** (Yannakoudakis et al., 2011) содержит тексты для изучающих английский язык на уровне B1–B2 в форме коротких эссе, писем или описаний. Каждый текст исправлен одним аннотатором. Датасет разделен на обучающую, валидационную и тестовую части.
2. **CoNLL-2014** (Ng et al., 2014) — часть корпуса *National University of Singapore Corpus of Learner English* (NUCLE; Dahlmeier et al., 2013). Этот датасет был создан как тестовый для соревнования CoNLL-2014 и содержит эссе обучающихся уровня C1. Разные версии CoNLL-2014 содержат исправления от 18 различных экспертов.
3. **Write & Improve (W&I) and LOCNESS (BEA-2019)** (Bryant et al., 2019) был разработан для соревнования BEA-2019 и включает эссе изучающих английских как L2 (уровней A1–C2) и студентов — носителей языка. Датасет разделен на тренировочную, валидационную и тестовую части, причем последняя размечена пятью экспертами.

Эти датасеты широко используются для обучения и оценки GEC-систем (Dale et al., 2012; Ng et al., 2014; Bryant et al., 2019).

Russian Error-Annotated Learner English Corpus (REALEC) (Vinogradova & Lyashevskaya, 2022) состоит из эссе студентов, изучающих английский язык, большинство из которых имеют уровни B1–B2. Ошибки были размечены четырьмя экспертами, которые использовали единое руководство по аннотации ошибок. Несмотря на новизну корпуса и тот факт, что он применялся лишь в одном крупном исследовании, посвященном GEC (Volodina et al., 2023), он особенно ценен для настоящей работы, поскольку содержит разметку дискурсивных ошибок.

Подробная сводка по датасетам приведена в Приложении A.

В случае FCE и BEA-2019 материал был взят только из тестовой части. Поскольку CoNLL-2014 является тестовым датасетом, а REALEC пока активно не использовался для обучения GEC-моделей, нет опасений, что модели могли обучиться на соответствующих примерах. Таким образом, риск перетекания тестовых данных в обучающие минимален.

Аннотация концентрированного датасета

Контекстно зависимые ошибки не обладают общими признаками, которые позволяли бы их легко автоматически извлекать, кроме того, они редко отдельно размечены в существующих датасетах и встречаются нечасто. Вследствие этого их извлечение из корпусов требует значительного объема ручной аннотации.

Для обеспечения соответствия аннотаций единому стандарту и сопоставимости предложений из разных источников мы унифицировали разметку всех датасетов, обработав исправления и автоматически приписав теги с помощью инструмента Error Annotation Toolkit (ERRANT; Bryant et al., 2017). Кроме того, мы сохранили оригинальную аннотацию дискурсивных ошибок из REALEC.

Для оптимизации процесса аннотации мы сосредоточились на нескольких типах ошибок, которые, как предполагалось, чаще проявляют контекстную зависимость, согласно Bryant et al. (2021). Мы выбрали для рассмотрения ошибки с тегами ERRANT: CONJ (союзы), DET (артикли), NOUN:INFL (склонение существительных), PRON (местоимения), PUNCT (пунктуация), VERB:SVA (согласование подлежащего и сказуемого), VERB:TENSE (времена глаголов), WO (порядок слов), а также с некоторыми тегами из корпуса REALEC, такими как Inappropriate_register (стилистические ошибки), Linking_device (дискурсивные связующие элементы) и Ref_device (анафорические выражения). Для большинства этих типов зависимость от информации в предшествующем и/или последующем контексте очевидна и демонстрируется примерами в Приложениях B и C. В качестве примера можно выделить несколько типов таких ошибок, использование определенного артикля при первом упоминании или неопределенного артикля при последующем упоминании (аннотируется как DET); использование предикатов в настоящем времени при наличии ссылки на конкретное время в прошлом (аннотируется как VERB:TENSE) и т. д.

Предварительно было размечено около 50 примеров для каждого типа. Для дальнейшего анализа были отобраны теги с наибольшей долей чувствительных к контексту ошибок в том числе теги ERRANT: PRON, PUNCT, VERB:TENSE и теги REALEC: Inappropriate_register, Linking_device и Ref_device. Затем для каждого из этих типов ошибок было размечено от 140 до 260 примеров. Число аннотированных примеров и доля контекстно зависимых ошибок каждого типа приведены в Приложении B. Описание менее частотных тегов представлено в Приложении C.

Для каждого предложения указывался изначальный тег, присвоенный инструментом аннотации ERRANT или аннотатором (для REALEC). Эксперты из команды авторов должны были изучить предложение в контексте и определить, необходимо ли учитывать информацию из других клауз или предложений для обнаружения и/или исправления ошибки. Каждый такой пример сопровождался расширенным описанием:

- требуется ли контекст из другой клаузы/предложения для обнаружения ошибки;
- требуется ли контекст из другой клаузы/предложения для исправления ошибки;
- тип контекста, необходимого для обнаружения или исправления ошибки: межклаузальный или межфразовый;
- расстояние в предложениях или клаузах (если контекст находится в том же предложении) от предложения/клаузы с ошибкой;

- направление, в котором находится контекст: слева, справа, с любой стороны или с обеих сторон от ошибочного предложения или клаузы;
- тип ошибки (см. Приложения B и C).

В результате мы обработали и аннотировали 3403 потенциально чувствительные к контексту ошибки, полученные из четырех корпусов текстов для учащихся, не являющихся носителями английского языка, и отобрали в общей сложности 1014 контекстно зависимых ошибок.

Согласованность аннотаторов

Для оценки валидности полученной разметки был проведен расчет согласованности между аннотаторами. Для этого случайным образом были отобраны 100 предложений, представляющих все исходные типы ошибок. Все предложения были размечены четырьмя экспертами, работавшими над полным датасетом. Этот набор примеров (далее называемый датасетом согласованности) использовался для расчета согласованности аннотаторов по полю «требуется ли контекст из другой клаузы/предложения для обнаружения ошибки».

Поскольку датасет согласованности не содержал пропущенных значений и аннотаторов было больше двух, мы использовали коэффициенты альфа Криппендорфа (Krippendorff, 2011), каппа Флейсса (Fleiss, 1971) и каппа Рэндольфа (Randolph, 2005; Warrens, 2010). Основной проблемой при расчете метрик стал сильный дисбаланс классов в датасете: доля контекстно зависимых ошибок была относительно мала по сравнению с общим числом ошибок. Это исказило оценку случайного совпадения аннотаций и привело к занижению показателей согласованности по обычно используемым коэффициентам (альфе Криппендорфа и каппе Флейсса).

Эффект резкого снижения альфы Криппендорфа при аннотации небольших и несбалансированных наборов данных (Marzi et al., 2024) проиллюстрирован в Таблице 1. В таблице представлено, как этот коэффициент работает на примере гипотетического датасета, который был аннотирован тремя группами экспертов. Первая группа показывает идеальную согласованность; во второй группе одна ошибка в аннотации; в третьей группе два эксперта сделали по одной ошибке. Можно заметить, что даже одна ошибка приводит к снижеению показателя до 0,429, а вторая ошибка обнуляет его, несмотря на то, что интуитивно согласованность аннотаторов остается относительно высокой.

Для компенсации этого эффекта мы рассчитали каппу Рэндольфа, которая менее чувствительна к дисбалансу классов. Кроме того, мы использовали собственную меру согласованности: каждую аннотацию из датасета согласованности сравнивали с аннотацией, включенной в основной датасет, и вычисляли процент совпадений. Затем рассчитывался средний процент для подмножества:

$$\frac{100}{n} \times \sum_{i=1}^{n} \frac{\sum_{j=1}^{m} [item_{i,j} = base_i]}{m},$$

где n — количество данных в датасете;
m — количество аннотаторов;
$item_{i,j}$ — аннотация j-го аннотатора для i-го элемента данных;
$base_i$ — аннотация из основного датасета, соответствующая i-му элементу данных.

Хотя этот метод нестандартен, он позволяет приблизительно оценить степень соответствия разметки датасета согласованности основному датасету и помогает соотнести другие метрики согласованности.

Поскольку традиционные показатели, такие как альфа Криппендорфа и каппа Флейсса, оказались довольно низкими, мы дополнительно рассчитали показатели согласованности для каждой отдельной категории ошибок, чтобы продемонстрировать, какие категории были наиболее и наименее надежными. Результаты для отдельных категорий ошибок, а также для всего датасета согласованности приведены в Таблице 2.

Все четыре метрики указывают на идеальную согласованность при значении 1 (или в нашем случае, 100, поскольку мы приводим результаты в процентах), но требуют разной интерпретации:

Таблица 1

Иллюстрация быстрой деградации альфы Криппендорфа на несбалансированных наборах данных

Ожидаемое значение[a]	Пример аннотации 1[b]			Пример аннотации 2[b]			Пример аннотации 3[b]		
	A1.1	A1.2	A1.3	A2.1	A2.2	A2.3	A3.1	A3.2	A3.3
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	1	1	2	1
1	1	1	1	1	1	1	1	1	1
K's αc	1,000			0,429			0,000		

Примечание: Правильное значение аннотации, ожидаемое для гипотетического датасета. Аннотации, выполненные тремя группами экспертов. Серые ячейки показывают случаи некорректной аннотации экспертом. Альфа Криппендорфа.

- Альфа Криппендорфа может принимать отрицательные (указывая на большее, чем случайное, расхождение в разметке разных экспертов) или положительные значения (но не превышающие 1). Согласованность выше 0,67 считается достаточно высокой для осторожных выводов, выше 0,8 — надежной. Несмотря на то что для нашего датасета согласованность невысока, из-за небольшого размера выборки показатель может быть занижен. С учетом этого можно предположить, что «реальная» согласованность составляет не менее 0,67.

- Каппа Флейсса и каппа Рэндольфа интерпретируются почти так же, как альфа Криппендорфа: согласованность выше 0,6 считается существенной, а выше 0,8 — практически идеальной. Для нашего датасета значения приближаются к порогу 0,6 для каппы Флейсса и превышают его для каппы Рэндольфа. Учитывая чувствительность метрики к размеру выборки, можно предположить, что реальная согласованность значительна.
- Авторская метрика, хотя и не имеет стандартной интерпретации, показывает высокие значения: в среднем 9 из 10 аннотаций совпадают с представленными в основном датасете.

Оценка контекстно зависимых ошибок

Процедура оценки имеет ключевое значение для контекстно зависимого исправления ошибок, поскольку ее стандартная реализация для задачи GEC систематически приводит к более низким показателям для длинных текстов независимо от их содержания.

Расчет оценки

Наиболее распространенной метрикой для оценки качества предсказаний является F_β-мера. В контексте задачи GEC чаще всего используется мера $F_{0,5}$ (Ng et al., 2014). Предполагается, что она лучше всего соответствует мнению экспертов о качестве исправления текста (Grundkiewicz et al., 2015; Napoles et al., 2015; Chollampatt & Ng, 2018).

F_β-оценка — это составная метрика, учитывающая истинно положительные предсказания (TP — случаи, когда модель сделала правильное исправление), ложно отрицательные (FN — случаи, когда модель не исправила ошибку, которую должна была исправить) и ложно положительные (FP — случаи, когда модель изменила текст, не содержащий ошибок). Точность (Precision) = TP / (TP + FP) и полнота (Recall) = TP / (TP + FN) рассчитываются как промежуточные шаги, а множитель $\beta = 0,5$ придает вдвое больший вес точности по сравнению с полнотой в результирующем значении. Простая интерпретация этой метрики такова: чем выше оценка, тем лучше работает соответствующая модель.

Наиболее современная реализация этого подхода представлена в инструменте ERRANT (Bryant et al., 2017), который включает следующие шаги:

(1.) Подготовка данных. Инструмент принимает исходный текст с ошибками и версии этого же текста, исправленные экспертами. Формируются наборы исправлений, которые необходимо применить к исходному тексту, чтобы получить версии экспертов.
(2.) Расчет **$F_{0,5}$-меры.** Для каждого набора правок, соответствующего исправлениям одного аннотатора, рассчитываются значения TP, FN и FP, по которым вычисляется $F_{0,5}$-мера.
(3.) Выбор ближайшего аннотатора. Из всех экспертных разметок выбирается вариант с наивысшей $F_{0,5}$-мерой, и соответствующие TP, FN и FP принимаются за эталонные для данного текста.
(4.) Итерация по текстам. Повторение шагов 1–3 для каждого текста (по умолчанию каждое предложение обрабатывается отдельно).
(5.) Финальная оценка. Значения TP, FN и FP, полученные на каждой итерации для ближайшего аннотатора, суммируются, чтобы рассчитать общую $F_{0,5}$-меру.

Таблица 2

Метрики согласованности аннотаторов

Тип ошибки	Альфа Криппендорфа	Каппа Флейсса	Каппа Рэндольфа	Авторская метрика	Доля ошибок, чувствительных к контексту [a]
Все типы	55,9	55,9	73,6	89,5	9/100
DET	N/A	N/A	N/A	100	0/18
Inappropriate register	0,0	-5,3	80,0	95,0	0/5
Linking device	60,4	58,3	60,0	90,0	2/5
PRON	35,0	34,0	56,9	82,4	0/17
PUNCT	53,2	52,6	55,0	81,2	2/20
Ref_device	54,8	52,4	60,0	90,0	2/5
VERB:TENSE	74,0	73,5	77,8	90,0	3/15
WO	-1,7	-3,4	86,7	96,7	0/15

Примечание. *[a] Средний процент аннотаций, соответствующих представленным в основном наборе данных.*

Нетривиальной особенностью алгоритма является возможность использовать несколько аннотаций и ориентироваться на наиболее близкий вариант исправления. Это демонстрирует многообразие языковых выражений и возможность различных способов коррекции одних и тех же ошибок. Как следствие, оценка по единственному варианту исправления без учета альтернатив недостаточна для объективного анализа эффективности модели. Более подробное обсуждение этой проблемы представлено в Bryant and Ng (2015).

Взаимосвязь длины текста и $F_{0,5}$-меры

Описанный алгоритм оценки и его подход к учету вариативности корректных исправлений существенно влияют на анализ контекстно зависимых ошибок.

В текстах на втором языке наблюдается высокая плотность ошибок (часто более одной ошибки на предложение). Кроме того, каждая ошибка вносит возможную вариативность, в связи с чем комбинаторика корректных исправлений может стать сложной. Некоторые корректные версии текста, сгенерированные GEC-системой, могут отсутствовать в эталонных аннотациях и быть несправедливо сочтены ошибочными.

Чтобы минимизировать этот эффект, тексты разбиваются на предложения: чем меньше текстовый фрагмент, предлагаемый для оценки, и чем меньше ошибок он содержит, тем полнее учитывается вариативность и тем точнее становится оценка. Если текстовые фрагменты достаточно малы, можно ожидать, что большое количество аннотаторов покроет все комбинации возможных исправлений внутри них.

Для демонстрации этого мы использовали ERRANT и оценивали модель BART (Katsumata & Komachi, 2020) на датасете CoNLL-2014. Мы рассчитали два показателя для одних и тех же выходных данных, предоставленных моделью, примененной на уровне полного текста. Первый показатель следовал обычному рабочему процессу ERRANT, включая разбиение датасета на предложения (следует заметить, что модель все равно применялась на уровне текста). Второй показатель отличался тем, что оценка проводилась для целых текстов. Результаты представлены в Таблице 3. Кроме того, мы рассчитали показатель для модели, примененной на уровне предложений, и установили, что модель хуже обрабатывает более длинные тексты (что подтверждает одну из основных гипотез исследования).

Таблица 3

Оценка модели BART для предложений и текстов

Применение модели	Измерение оценок	FP	FN	TP	Precision	Recall	$F_{0,5}$
По тексту	по предложениям	507	1589	1111	68,67	41,15	60,56
	по тексту	620	2131	1012	62,01	32,2	52,32
По предложениям	по предложениям	216	978	1367	86,36	58,29	78,77

Два значения, полученные для одного и того же предсказания, различаются: как предполагалось, оценка становится ниже при рассмотрении более длинных единиц. Примечательно, что при условии полной корректности аннотаций именно более высокая оценка лучше характеризует производительность модели, что делает разбиение текстов на предложения разумным подходом к оценке.

Однако это решение проблематично с точки зрения контекстно зависимых ошибок, исправление которых по определению требует обработки более длинных контекстов. В некоторых случаях контекстно зависимые ошибки расположены внутри одного предложения, и, хотя для их исправления требуется контекст, это не влияет на оценку. Однако это не всегда так.

Во-первых, существуют ошибки, расположенные на стыке предложений. Самый простой пример — ошибка пунктуации, такая как замена точки на запятую или наоборот, но возможны и более сложные случаи.

Во-вторых, некоторые ошибки зависят друг от друга: бывает, что две ошибки должны быть исправлены согласованно. Например, появление заглавной буквы в начале сегмента, оно сопутствует замене запятой на точку. Другой часто встречающийся пример — это последовательности глаголов, использованных в неправильном времени, когда вся последовательность зависит от широкого контекста слева. В таких случаях раздельное рассмотрение предложений проблематично: аннотаторы могут по-разному интерпретировать временную форму для всей последовательности (например, один выбирает Past Simple, а другой — Present Simple). Если последовательность разделена на предложения, переключение между Past и Present Simple будет ошибочно оценено моделью как корректное.

Чтобы учесть эти проблемы, оценка концентрированного датасета производится следующим образом:

(1.) Для баланса между необходимостью оценивать минимальные текстовые фрагменты и возможностью некорректного учета контекстно зависимых ошибок тексты разделяются на минимальные блоки, в которых предложения с взаимозависимыми ошибками не разделены. То есть, если ошибка произошла на границе предложения или ее исправление требовало объединения двух или более предложений, все вовлеченные предложения включались учитывались при оценке как единое целое.

(2.) Мы оцениваем только аннотированные контекстно зависимые ошибки, что позволяет минимизировать искажения, вызванные увеличением охватываемых контекстов.
(3.) Мы сохраняем только одну аннотацию для контекстно зависимых ошибок, учитывая, что в процессе ручной разметки не было выявлено крупномасштабной вариативности (вариации исправлений были найдены для ошибок в местоимениях, но таких исключений очень мало).

Организация эксперимента

Для тестирования концентрированного датасета мы использовали две современные GEC-модели: BART (large; Katsumata & Komachi, 2020) и T5 (base; Rothe et al., 2021). Мы выбрали эти две модели вместо более новых GEC-систем (Zhou et al., 2023), поскольку последние обычно используют одну или несколько LLM из стандартного набора, добавляя дополнительные компоненты пред- или постобработки. Учитывая, что общие результаты сопоставимы, мы выбрали менее сложные конструкции для получения более интерпретируемого результата.

РЕЗУЛЬТАТЫ

Концентрированный датасет контекстно зависимых ошибок

Одним из практических результатов исследования стало создание концентрированного датасета с более высокой долей контекстно зависимых ошибок.[1] Датасет содержит 1014 контекстно зависимых ошибок с дополнительной разметкой.

В Таблицах 4–6 представлена общая информация о датасете: распределение типов ошибок, представленность исходных датасетов в концентрированном датасете, а также тип контекста, необходимого для исправления ошибки.

Как показано в Таблице 4, концентрированный датасет включает 5 основных типов контекстно зависимых ошибок: ошибки в местоимениях (PRON), пунктуации (PUNCT), референциальных средствах (REF), времени глагола (VERB:TENSE) и коннекторах (LINK). Другие типы либо имеют низкую долю контекстно зависимых ошибок и не были подробно аннотированы (например, WO — порядок слов), либо возникли в результате ручной корректировки неточной автоматической атрибуции тегов ERRANT.

Таблица 6 демонстрирует, что в большинстве случаев именно левый (предшествующий единице) контекст определяет способ исправления ошибки. Чаще всего для исправления достаточно одного предложения, но представлен и более широкий диапазон дистанций. Правый (последующий) контекст требуется примерно в одном из 20 случаев, и почти всегда это соседнее предложение. В редких случаях достаточно либо левого, либо правого контекста[2], и лишь в очень немногих случаях необходимы оба контекста.

Оценка GEC-моделей на концентрированном датасете

Для тестирования выбранного подхода мы измерили производительность современных моделей BART и T5 на созданном датасете. Результаты оценки представлены в Таблице 7.

Прежде чем обсуждать закономерности в данных, необходимо прокомментировать особенность метода измерения, которая существенно влияет на результаты. Число ложноположительных результатов (False Positives) в таблице равно нулю для каждой строки. Следовательно, для каждой строки точность (Precision) равна 100 %. Это непосредственно следует из процедуры измерения, описанной выше в разделе «Материалы и методы»: мы оцениваем только контекстно зависимые ошибки и не рассматриваем другие типы ошибок, которым автоматически присваиваются ложноположительные результаты. Для корректной оценки реального числа контекстно зависимых ложноположительных результатов потребовалось бы вручную обработать все такие случаи и определить, зависят ли они от широкого контекста.

В результате показатели $F_{0,5}$ в таблице следует рассматривать как максимальные значения оценки. Полное отсутствие ложноположительных результатов в выдаче даже наиболее эффективной модели крайне маловероятно, поэтому фактические значения $F_{0.5}$-меры ниже. Учитывая, что коэффициент 0,5 в F-мере придает вдвое больший вес точности по сравнению с полнотой, значения в Таблице 7 гораздо более оптимистичны, чем должны быть, и правильнее рассматривать истинно положительные и ложноотрицательные результаты, а также полноту.

Даже с учетом этих оговорок оценки в Таблице 7 явно ниже, чем оценки для ошибок, не чувствительных к широкому контексту. Общие показатели BART составляют 40,39 для контекстно зависимых ошибок против 78,04 для не контекстно зависимых (последние измерены на датасете CoNLL-2014); для T5 эти показатели составляют 10,85 и 74,38 соответственно.

[1] https://huggingface.co/datasets/startc/doc-gec

[2] Если информация, необходимая для исправления, содержится как в левом, так и в правом контексте, но расположена ближе к одной из сторон, учитывается только ближайший контекст. Например, если клаузу можно исправить на основе предыдущего предложения или предложения, отстоящего от нее на четыре предложения правее, мы учитываем только левый контекст.

Таблица 4

Типы ошибок в концентрированном наборе данных

Тип ошибки	Кол-во примеров
PRON	259
PUNCT	202
REF	201
VERB:TENSE	171
LINK	140
DET	19
VERB:MODAL	8
Другие типы	14
Итого	1014

Таблица 5

Представленность исходных датасетов в концентрированном датасете

Датасет	Кол-во использованных примеров
REALEC	633
BEA-2019	218
FCE	135
CoNLL-2014	28
Итого	1014

Таблица 6

Контекст, необходимый для исправления ошибки

Тип контекста	Число предложений, необходимых для выявления и исправления	Кол-во ошибок
Левый	1	810
	2	64
	3	22
	4	17
	>4	21
Левый, итого		934
Правый	1	59
	2	1
	>4	1
Правый, итого		61
Левый или правый	1 слева, 1 справа	16
Левый и правый	1 слева, 1 справа	2
	3 слева, 1 справа	1
Левый и правый, итого		3
Итого		1014

Если рассматривать полноту, низкая производительность моделей на контекстно зависимых ошибках становится еще более заметной. В случае PUNCT как наиболее хорошо обрабатываемого типа ошибок и BART как лучшей модели лишь 27,23 % ошибок исправляются корректно. Остальные результаты еще ниже, и большинство значений (все для T5; REF и LINK для BART) близки к уровню шума с точки зрения количества корректных исправлений, выполненных моделью. В то же время датасет позволяет четко увидеть разницу в качестве работы моделей: BART стабильно показывает более высокие результаты, чем T5. Это подтверждает, что низкие результаты, полученные на концентрированном

Таблица 7

Оценка моделей BART и T5 на концентрированном датасете

Тип ошибки	BART						T5					
	FP	FN	TP	Prec	Rec	$F_{0,5}$	FP	FN	TP	Prec	Rec	$F_{0,5}$
Все типы	0	893	121	100	11,93	40,39	0	986	24	100	2,38	10,85
PUNCT	0	147	55	100	27,23	65,17	0	192	10	100	4,95	20,66
VERB:TENSE	0	142	29	100	16,96	50,52	0	165	4	100	2,37	10,81
PRON	0	239	20	100	7,72	29,5	0	197	4	100	1,99	9,22
REF	0	192	9	100	4,48	18,99	0	256	3	100	1,16	5,54
LINK	0	138	2	100	1,43	6,76	0	140	0	100	0	0

датасете, не сводятся к его внутренним особенностям, а выявляют недостатки GEC-систем в отношении выбранных типов ошибок.

Наконец, интересно отметить закономерность в различии метрик для разных типов ошибок. Наивысшие оценки наблюдаются для пунктуации, которая представляет собой искусственно регулируемую конструкцию над письменной системой языка, и времени глаголов, которое является единственным чисто грамматическим типом ошибок в выборке. Местоимения как анафорические средства в большей степени относятся к уровню дискурса, хотя обычно рассматриваются как часть грамматики, тогда как тип Referential_device включает лексически закодированные (и менее связанные с грамматикой) анафорические средства. Наконец, связующие конструкции относятся исключительно к сфере дискурса. Таким образом, можно утверждать, что качество исправления ошибок снижается по мере смещения типа ошибки от грамматики к дискурсу.

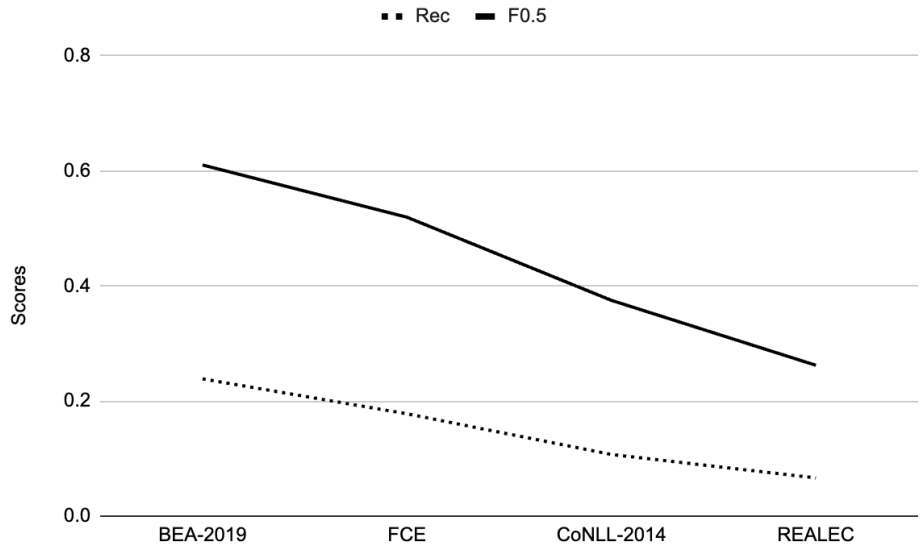
ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Анализируя схему оценивания для задачи GEC с учетом контекстно зависимых ошибок, можно сделать несколько наблюдений. Во-первых, полученные результаты показывают, что концентрированный датасет помогает более реалистично оценить производительность моделей. В борьбе с инфляцией метрик, наблюдаемой на традиционно используемых в GEC датасетах, концентрированные датасеты могут служить дополнительными индикаторами недостатков моделей наряду с покатегориальной оценкой (Bryant et al., 2017).

Во-вторых, можно заметить, что метрики значительно различаются между четырьмя подмножествами, стратифицированными по источнику данных. Как показано на Рисунке 1, $F_{0,5}$-мера варьируется от 0,61 для BEA-2019 до 0,26 для REALEC, а полнота — от 0,24 для BEA-2019 до 0,07 для REALEC. Эти результаты согласуются с другими сравнительными ис-

Рисунок 1

Результаты оценки на подкорпусах BEA-2019, FCE, CoNLL-2014 и REALEC (модель BART)



следованиями GEC (Zhang et al., 2023; Volodina et al., 2023 и др.), подтверждая, что наблюдаемая вариативность может быть обусловлена множеством факторов, включая уровень владения языком, стиль текста, тип письменного задания, длину текста и предложения, стратегию аннотирования и соответствующие различия в распределении тегов ошибок. Однако важно отметить, что снижение производительности между подмножествами в концентрированных данных более выражено по сравнению с результатами, наблюдаемыми на не концентрированных датасетах.

В-третьих, хотя концентрированный датасет относительно мал для окончательных выводов, мы обнаружили взаимосвязь между типами ошибок и объемом контекста, необходимого для их обнаружения и исправления. Эта информация отражена в датасете как количество единиц контекста (предложений или клауз), необходимых для корректного нахождения и исправления ошибки. Например, подавляющее большинство ошибок VERB:TENSE требуют не более одной клаузы (см. пример (3)), в то время как ошибки, помеченные как LINK, обычно связаны с одним или несколькими предложениями в левом или правом контексте (см. пример (4)):.

(3) When I was little I **had**→[∅] tried a lot of sports...

(4) From 2000 the percentage of elderly people in Sweden began to rise to 20 per cent. **Moreover**→**[Contrary to that]**, from 2000 the percentage in the USA was at the same level of 14 per cent.

Дальнейшие исследования могут быть направлены на эксперименты с различными архитектурами и методами GEC для понимания вариативности метрик между датасетами и роли доступного контекста в производительности моделей.

Хотя различия в значениях $F_{0,5}$ для концентрированных и обычных датасетов очевидны, вопрос применимости этой метрики в задаче GEC остается открытым. С развитием методов инструктирования (prompting) генеративных моделей все чаще сообщается о равенстве или даже превосходстве полноты над точностью. В этой связи Zeng et al. (2024) предлагают использовать F_1 и F_2-меры как более репрезентативные метрики оценки результатов GEC. Как было показано, значения $F_{0,5}$, точности и полноты, рассчитанные для одной и той же модели при обработке целых текстов и отдельных предложений (см. Таблицу 3), не всегда прямо соответствуют друг другу. Для разработки эталонного подхода к оценке результатов для задачи GEC с учетом влияния широкого контекста необходима гармонизация перечисленных метрик.

Ограничения исследования

При составлении концентрированного датасета необходимо оценивать природу и ключевые свойства используемых корпусов. Дальнейшие исследования могут быть направлены на увеличение объема датасета, балансировку примеров по уровню владения языком и типам ошибок, а также привлечение большего числа экспертов для обеспечения надежности разметки.

Еще одно ограничение нашего подхода заключается в том, что представленный датасет является лишь первым шагом к более детальному изучению процедур отбора данных, методов оценки и обучения моделей. В нашей работе использовались только готовые модели. Очевидно, что для лучшего понимания успешности исправления ошибок, зависящих от широкого контекста, для разных типов ошибок необходимы дальнейшие эксперименты с обучением моделей на концентрированных (обучающих) датасетах.

ЗАКЛЮЧЕНИЕ

В настоящем исследовании мы предложили использование концентрированного датасета с высокой долей контекстно зависимых ошибок как способ решения проблемы ограничения разрешающей способности методов оценки для задачи GEC. Эта проблема возникает потому, что метрики, обычно используемые для оценки GEC-моделей, могут переоценивать качество их работы, хотя определенные типы ошибок часто остаются незамеченными. Путем ручной разметки примеров различных типов ошибок (связанных с пунктуацией, временами глаголов, артиклями, местоимениями, референциальными средствами и связующими конструкциями) мы создали датасет, содержащий 1014 ошибок для выявления и/или исправления которых требуется широкий контекст. Мы протестировали на этом датасете две модели GEC и показали, что их производительность значительно ниже на концентрированном датасете по сравнению с не концентрированным. Это подтверждает, что GEC-системы все еще нуждаются в значительных улучшениях, и подчеркивает потенциал концентрированных датасетов как инструмента для обучения и оценки.

На основании проведенного анализа производительности двух моделей на разных типах ошибок мы предполагаем, что сложность коррекции возрастает по мере перехода от грамматики в область дискурса. Например, ошибки в пунктуации и временах глаголов исправляются более успешно, чем ошибки, связанные с референциальными средствами и коннекторами.

Настоящая работа демонстрирует потенциал использования концентрированных датасетов с высокой долей контекстно зависимых ошибок для дальнейшего совершенствования GEC-систем и повышения их применимости для реальных задач. В качестве практического вклада мы публикуем концентрированный датасет в открытом доступе[3].

[3] https://huggingface.co/datasets/startc/doc-gec

БЛАГОДАРНОСТИ

Данная статья является результатом исследовательского проекта, выполненного в рамках Программы фундаментальных исследований Национального исследовательского университета «Высшая школа экономики» (НИУ ВШЭ).

КОНФЛИКТ ИНТЕРЕСОВ

Авторы заявляют об отсутствии конфликта интересов.

ВКЛАД АВТОРОВ

Владимир Старченко: концептуализация; сбор данных; разметка данных (автоматизация); исследование; методология; тестирование моделей; управление проектом; статистический анализ; научное руководство; написание исходного текста; редактирование текста.

Елизавета Клыкова: концептуализация; разметка данных; исследование; методология; управление проектом; написание исходного текста; редактирование текста.

Анастасия Шаврина: концептуализация; разметка данных; исследование; методология; редактирование текста.

Ольга Виноградова: концептуализация; разметка данных; исследование; методология; написание исходного текста.

Ольга Ляшевская: концептуализация; разметка данных; исследование; научное руководство; написание исходного текста; методология; управление проектом; редактирование текста.

Дарья Харламова: концептуализация; разметка данных; исследование; методология; ресурсы; написание исходного текста; редактирование текста.

Алексей Старченко: концептуализация; исследование; методология; управление проектом; статистический анализ; научное руководство; написание исходного текста; редактирование текста.

ЛИТЕРАТУРА

Bentley, J. (1985). Programming pearls: A spelling checker. *Communications of the ACM*, *28*(5), 456–462. https://doi.org/10.1145/3532.315102

Brockett, C., Dolan, B., & Gamon, M. (2006). Correcting ESL errors using phrasal SMT techniques. *21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL* (pp. 249–256). Association for Computational Linguistics. http://dx.doi.org/10.3115/1220175.1220207

Bryant, C., Felice, M., Andersen, Ø. E., & Briscoe, T. (2019). The BEA-2019 shared task on grammatical error correction. *Proceedings of the fourteenth workshop on innovative use of NLP for building educational applications* (pp. 52–75). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W19-4406

Bryant, C., Felice, M., & Briscoe, T. (2017). Automatic annotation and evaluation of error types for grammatical error correction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long Papers, pp. 793–805). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P17-1074

Bryant, C., Yuan, Z., Qorib, M. R., Cao, H., Ng, H. T., & Briscoe, T. (2023). Grammatical error correction: A survey of the state of the art. *Computational Linguistics*, *49*(3), 643–701. http://dx.doi.org/10.1162/coli_a_00478

Bryant, C., & Ng, H. T. (2015). How far are we from fully automatic high quality grammatical error correction? *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (vol. 1: Long Papers, pp. 697–707). Association for Computational Linguistics. http://dx.doi.org/10.3115/v1/P15-1068

Burstein, J., Chodorow, M., & Leacock, C. (2003). Criterion℠ online essay evaluation: An application for automated evaluation of student essays. *Proceedings of the Fifteenth Conference on Innovative Applications of Artificial Intelligence* (pp. 3–10). American Association for Artificial Intelligence.

Cargill, T. A. (1980). The design of a spelling checker's user interface. *ACM SIGOA Newsletter*, *1*(3), 3–4. https://doi.org/10.1145/1017923.1017924

Chollampatt, S., & Ng, H. T. (2018). A multilayer convolutional encoder-decoder neural network for grammatical error correction. *Proceedings of the AAAI conference on artificial intelligence* (vol. 32(1), pp. 5755–5762). Association for the Advancement of Artificial Intelligence. http://dx.doi.org/10.1609/aaai.v32i1.12069

Chollampatt, S., Wang, W., & Ng, H. T. (2019). Cross-sentence grammatical error correction. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 435–445). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/P19-1042

Dahlmeier, D., & Ng, H. T. (2011). Grammatical error correction with alternating structure optimization. *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies* (pp. 915–923). Association for Computational Linguistics.

Dahlmeier, D., Ng, H. T., & Wu, S. M. (2013). Building a large annotated corpus of learner English: The NUS corpus of learner English. *Proceedings of the eighth workshop on innovative use of NLP for building educational applications* (pp. 22–31). Association for Computational Linguistics.

Dale, R., Anisimoff, I., & Narroway, G. (2012). HOO 2012: A report on the preposition and determiner error correction shared task. *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP* (pp. 54–62). Association for Computational Linguistics.

Du, Z., & Hashimoto, K. (2023). Sentence-level revision with neural reinforcement learning. *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)* (pp. 202–209). Association for Computational Linguistics.

Grundkiewicz, R., Junczys-Dowmunt, M., & Gillian, E. (2015). Human evaluation of grammatical error correction systems. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (pp. 461–470). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D15-1052

Edunov, S., Ott, M., Auli, M., & Grangier, D. (2018). Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 489–500). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/D18-1045

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619

Jensen, K., Heidorn, G., Miller, L., & Ravin, Y. (1993). Parse fitting and prose fixing. *Natural Language Processing: The PLNLP Approach* (pp. 53–64). Springer. https://doi.org/10.1007/978-1-4615-3170-8_5

Katsumata, S., & Komachi, M. (2020). Stronger baselines for grammatical error correction using a pretrained encoder-decoder model. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing* (pp. 827–832). Association for Computational Linguistics.

Krippendorff, K. (2011). Computing Krippendorff's Alpha-Reliability. https://repository.upenn.edu/asc_papers/43

Kwasny, S. C., & Sondheimer, N. K. (1981). Relaxation techniques for parsing grammatically ill-formed input in natural language understanding systems. *American Journal of Computational Linguistics*, *7*(2), 99–108.

Lee, J. S. (2004). Automatic article restoration. *Proceedings of the Student Research Workshop at HLT-NAACL 2004* (pp. 31–36). Association for Computational Linguistics.

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. (2020). CrowS-Pairs: A challenge dataset for measuring social biases in masked language models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1953–1967). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.emnlp-main.154

Omelianchuk, K., Atrasevych, V., Chernodub, A., & Skurzhanskyi, O. (2020). GECToR–Grammatical Error Correction: Tag, not Rewrite. *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 163–170). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.bea-1.16

Leacock, C., Gamon, M., & Brockett, C. (2009). User input and interactions on Microsoft Research ESL assistant. *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 73–81). Association for Computational Linguistics.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2014). *Automated grammatical error detection for language learners* (2nd ed.). Morgan & Claypool Publishers. https://doi.org/10.1007/978-3-031-02153-4

Li, W., & Wang, H. (2024). Detection-correction structure via general language model for grammatical error correction. *arXiv preprint arXiv:2405.17804*. http://dx.doi.org/10.48550/arXiv.2405.17804

Marzi, G., Balzano, M., & Marchiori, D. (2024). K-Alpha Calculator–Krippendorff's Alpha Calculator: A user-friendly tool for computing Krippendorff's Alpha inter-rater reliability coefficient. *Methods X, 12*, *102545*. https://doi.org/10.1016/j.mex.2023.102545

Napoles, C., Sakaguchi, K., Post, M., & Tetreault, J. (2015). Ground truth for grammatical error correction metrics. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Vol. 2: Short Papers, pp. 588–593). Association for Computational Linguistics. http://dx.doi.org/10.3115/v1/P15-2097

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014). The CoNLL-2014 shared task on grammatical error correction. In *Proceedings of the eighteenth conference on computational natural language learning: Shared task* (pp. 1–14). Association for Computational Linguistics. http://dx.doi.org/10.3115/v1/W14-1701

Qorib, M. R., & Ng, H. T. (2022). Grammatical error correction: Are we there yet? In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 2794–2800). International Committee on Computational Linguistics.

Randolph, J. J. (2005). Free-marginal multirater kappa (multirater K[free]): An alternative to fleiss' fixed-marginal multirater kappa. *Presented at the Joensuu Learning and Instruction Symposium 2005* (October 14–15, 2005). http://files.eric.ed.gov/fulltext/ED490661.pdf

Rothe, S., Mallinson, J., Malmi, E., Krause, S., & Severyn, A. (2021). A simple recipe for multilingual grammatical error correction. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (vol. 2: Short Papers, pp. 702–707). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2021.acl-short.89

Rozovskaya, A., & Roth, D. (2010). Training paradigms for correcting errors in grammar and usage. *Human language technologies: The 2010 annual conference of the north american chapter of the association for computational linguistics* (pp. 154–162). Association for Computational Linguistics.

Rozovskaya, A. & Roth, D., (2021). How good (really) are grammatical error correction systems? *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (pp. 2686–2698). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.231

Sakaguchi, K., Napoles, C., Post, M., & Tetreault, J. (2016). Reassessing the goals of grammatical error correction: Fluency instead of grammaticality. *Transactions of the Association for Computational Linguistics*, *4*, 169–182. http://dx.doi.org/10.18653/v1/P18-1020

Starchenko, V. M., & Starchenko, A. M. (2023). Here we go again: modern GEC models need help with spelling. *Proceedings of ISP RAS*, *35*(5), 215–228. http://dx.doi.org/10.15514/ISPRAS-2022-35(5)-14

Starchenko, V. M. (2024). No need to get wasteful: The way to train a lightweight competitive spelling checker. *Computación y Sistemas*, 28(3), 1–12. https://doi.org/10.13053/CyS-28-4-5068

Vinogradova, O., & Lyashevskaya, O. (2022). Review of practices of collecting and annotating texts in the learner corpus REALEC. *International Conference on Text, Speech, and Dialogue* (pp. 77–88). Springer International Publishing. http://dx.doi.org/10.1007/978-3-031-16270-1_7

Volodina, E., Bryant, C., Caines, A., De Clercq, O., Frey, J., Ershova, E., Rosen, A., & Vinogradova, O. (2023). MultiGED-2023 shared task at NLP4CALL: Multilingual grammatical error detection. *Linköping Electronic Conference Proceedings* (pp. 1–16). LiU Electronic Press. https://doi.org/10.3384/ ecp197001

Wang, C., Li, R., & Lin, H. (2017). Deep context model for grammatical error correction. *SLaTE* (pp. 167–171). International Speech Communication Association. http://dx.doi.org/10.21437/SLaTE.2017-29

Wang, Y., Xia, Y., He, T., Tian, F., Qin, T., Zhai, C., & Liu, T. Y. (2019). Multi-agent dual learning. *Proceedings of the International Conference on Learning Representations* (ICLR). International Conference on Learning Representations.

Wang, Y., Wang, Y., Dang, K., Liu, J., & Liu, Z. (2021). A comprehensive survey of grammatical error correction. *ACM Transactions on Intelligent Systems and Technology*, *12*(5), 1–51. http://dx.doi.org/10.1145/3474840

Warrens, M. J. (2010). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification, 4*(4), 271–286. https://doi.org/10.1007/s11634-010-0073-4

Xie, Z., Avati, A., Arivazhagan, N., Jurafsky, D., & Ng, A. Y. (2016). Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727*. http://dx.doi.org/10.48550/arXiv.1603.09727

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 180–189). Association for Computational Linguistics.

Yuan, Z., & Bryant, C. (2021). Document-level grammatical error correction. *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 75–84). Association for Computational Linguistics.

Yuan, Z., & Felice, M. (2013). Constrained grammatical error correction using statistical machine translation. *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task* (pp. 52–61). Association for Computational Linguistics.

Yuan, Z., Briscoe, T., & Felice, M. (2016). Candidate re-ranking for SMT-based grammatical error correction. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 256–266). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/W16-0530

Zeng, M., Kuang, J., Qiu, M., Song, J. and Park, J. (2024). Evaluating prompting strategies for grammatical error correction based on language proficiency. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (LREC-COLING 2024) (pp. 6426–6430). ELRA and ICCL. https://doi.org/10.48550/arXiv.2402.15930

Zhang, Y., Zhang, B., Li, Z., Bao, Z., Li, C., & Zhang, M. (2022). SynGEC: Syntax-enhanced grammatical error correction with a tailored GEC-oriented parser. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 2518–2531). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2022.emnlp-main.162

Zhao, J., Fang, M., Pan, S., Yin, W., & Pechenizkiy, M. (2023). GPTBIAS: A comprehensive framework for evaluating bias in large language models. *arXiv preprint arXiv:2312.06315*. http://dx.doi.org/10.48550/arXiv.2312.06315

Zhou, H., Liu, Y., Li, Z., Zhang, M., Zhang, B., Li, C., Zhang, J., & Huang, F. (2023). Improving Seq2Seq grammatical error correction via decoding interventions. *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 7393–7405). Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2023.findings-emnlp.495

ПРИЛОЖЕНИЕ А

Общая информация об исходных датасетов, использованных для формирования концентрированного датасета

Датасет	Объем (в токенах)	Кол-во наборов аннотаций на документ [a]	Кол-во типов ошибок	Уровень владения языком
FCE (тестовая часть)	41,9 тыс.	1	71	B1-B2
CoNLL-2014	30,1 тыс.	2–18	28	C1
BEA-2019 (тестовая часть)	85,7 тыс.	5	55	A1-Native
REALEC	1550,6 тыс.	1	48	B1-B2

Примечание. *[a] Количество наборов аннотаций (от разных экспертов), предоставленных для каждого документа.*

ПРИЛОЖЕНИЕ В

Теги ошибок, используемые в датасете, и доля контекстно зависимых ошибок

Исходный тег [a]	Новый тег [b]	Доля контекстно зависимых ошибок [c]	Описание	Пример [d]
Linking_device	LINK	59,05%	Неправильно использовано или отсутствует связующее средство (коннектор)	Secondly, the majority of the population will use other kinds of public transport, for example, trains, cars, or ships. So→However, we cannot say that these types of transport harm our environment less than planes do.
Ref_ device	REF	50,83%	Неправильно использовано референциальное средство	We should not create barriers for ambitious people and accept persons→those who don't have interest in education just because of sex equality.
VERB: TENSE	VERB: TENSE	45,35%	Неправильно выбрано время глагола	When I was small, we lived in the country. I remembered→remember, we used to have oil lamps which used a cotton string dipping in the oil in the small bottle and made it burn the tip of the cotton string to give us light during the night.
PUNCT	PUNCT	37,61%	Неправильно использован знак пунктуации	In Sweden the level fell from 84% to 15%, a similar situation was in France. The→: the level changed from 90% to 50%.
PRON	PRON	36,72%	Неправильно использовано или отсутствует личное местоимение	Also, he is very funny and I laugh a lot with him. Both→We both like to travel around the world and to do some sports, for example, tennis, running or trekking.
Inappropriate_ register	REF, PRON[4]	15,50%	Ошибки, связанные со стилем и уместностью	When a child begins learning, for example, English in primary school, he→they get the necessary basis for further studying. (Tagged as PRON) Unfortunately, watching sports doesn't teach us→viewers anything and people don't get any information about the surrounding world from it. (Tagged as REF)
DET	DET	9,45%	Неправильно использован или отсутствует артикль	This situation creates a lot of pollution for ∅→the environment, so we have to be more concerned about the planet's health.

Примечания. *[a] Тег, используемый в исходном датасете. [b] Тег, используемый в концентрированном датасете. Доля контекстно зависимых ошибок среди всех ошибок с данным тегом. [d] Для наглядности все остальные ошибки в примерах были исправлены в соответствии с правками, предложенными аннотаторами исходных датасетов.*

[4] В процессе аннотирования мы пришли к выводу, что другие теги (такие как PRON или REF) более подходят для контекстно зависимых примеров, помеченных как Inappropriate_register в REALEC.

ПРИЛОЖЕНИЕ С

Прочие теги, используемые в концентрированном датасете

Тег	Описание	Пример
LEX	Ошибка в выборе лексической единицы	Also, it is a good way to get some positive emotions. All of this→Watching sports can even promote future productivity at work.
NOUN:NUM	Существительное использовано в неправильном числе	By the way, there is an opposite tendency with young people, their number→numbers are the largest at the science courses and the smallest in the sports and health courses. Additionally, students of the health and sports course→courses are mostly middle-aged.
SPELL	Орфографическая ошибка	To sum up, both characteristics are important in our life. We need to know how to operate with once→ones we were born with and know how to develop knowledge gained from our experience to have a successful life and reach goals we set for ourselves.
SYN	Неправильный выбор или ошибочное изменение синтаксической структуры	Although the grandparents are in most cases ready to help, they can not transfer the values of the new world to the kids, and their→this results in the wrong choice of paths of life for the grown-up adults in future.
VERB:MODAL	Модальный глагол отсутствует, избыточен или использован неправильно	In addition, to decrease the risk of negative comments or posts, Facebook and Twitter would→should improve their futures by solving the personal privacy problem.
VERB:SVA	Ошибки согласования подлежащего и сказуемого	Today, public transport still play→plays an important role in the transport system and it will keep on doing so in the future.
WO	Ошибки порядка слов (например, отсутствие инверсии в требуемых контекстах)	But when I was a teenager, I began to experience situations that I did not like, for instance, girls said to me bad things→bad things to me or they talked unkindly about me.