Методы автоматической оценки качества генерации текста больших языковых моделей: сравнительный анализ метрик и подходов

Д. Л. Тетеревенков

Финансовый университет при Правительстве РФ

249453@edu.fa.ru

Аннотация. **В статье представлен обзор современных автоматических метрик для оценки качества генерации текста большими языковыми моделями (LLM). Рассмотрены основные подходы, включая метрики на основе n-грамм, семантические и контекстуальные методы. Проведен сравнительный анализ метрик по ключевым критериям: точность, воспроизводимость, зависимость от домена и вычислительная сложность. Особое внимание уделено необходимости адаптации метрик к конкретным задачам и кейсам.**

Ключевые слова: большие языковые модели; автоматическая оценка текста; BLEU; ROUGE; BERTScore; сравнительный анализ

I. Введение

Развитие больших языковых моделей (LLM) требует надежных методов автоматической оценки качества генерируемого текста. Традиционные метрики, такие как BLEU и ROUGE, фокусируются на поверхностном совпадении с эталоном, однако современные подходы учитывают семантику, когерентность и контекст. В докладе проводится сравнительный анализ существующих метрик, выделяются их сильные и слабые стороны, а также предлагается методология для выбора оптимальных решений в зависимости от задачи.

A. Существующие автоматические метрики оценки генерации больших языковых моделей

- BLEU (Bilingual Evaluation Understudy)

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{4} w_n \ln p_n\right)$$

Метрика BLEU сравнивает частоту n-грамм (1–4 слов) в сгенерированном тексте с эталоном, учитывая модифицированную точность — например, если слово «кот» встречается в эталоне 2 раза, а в генерации 3 раза, засчитывается только 2 совпадения — и вводит штраф за краткость (brevity penalty), чтобы избежать завышенных оценок для слишком коротких текстов.

- ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

Метрика ROUGE включает две основные версии: ROUGE-N, которая анализирует совпадение n-грамм между сгенерированным текстом и эталоном (аналогично BLEU, но с фокусом на полноту), и ROUGE-L, сравнивающую самую длинную общую подпоследовательность (LCS). Например, для эталонного текста «Кот спит на ковре» и сгенерированного варианта «На ковре спит кот» LCS составляет последовательность «спит кот», что дает значение ROUGE-L = 2/4, где числитель — длина LCS, а знаменатель — длина эталона. Преимущество ROUGE-L заключается в учете порядка слов в пределах подпоследовательности, что повышает точность оценки связности текста.

- METEOR (Metric for Evaluation of Translation with Explicit ORdering)

Метрика, которая сопоставляет слова из генерации и эталона, учитывая синонимы (с использованием лингвистической базы WordNet) и морфологические вариации (например, преобразуя «бежал» в базовую форму «бежать»). Кроме того, METEOR вводит штраф за несовпадение порядка слов, что позволяет более точно оценивать структурную целостность текста по сравнению с метриками, ориентированными только на лексические совпадения.

- BERTScore

$$BERTScore = \frac{1}{N}\sum_{i=1}^{N} \max_j \cos\left(e_i^{ref}, e_j^{gen}\right)$$

Метрика BERTScore основана на использовании предобученной модели BERT. Её алгоритм работает следующим образом: сначала проводится токенизация эталонного и сгенерированного текста, затем для каждого токена получают контекстуальные эмбеддинги с помощью модели BERT, после чего вычисляется косинусное сходство между эмбеддингами соответствующих слов из эталона и генерации. На последнем этапе полученные значения усредняются по всем токенам.

- MoverScore

Метрика, основанная на расстоянии Earth Mover's Distance (EMD), которое измеряет «стоимость» переноса смысловых единиц между текстами. Алгоритм включает два основных действия: преобразование эталона и генерации в эмбеддинги (аналогично BERTScore) и расчёт минимальной «работы», необходимой для преобразования распределения смысловых единиц генерации в распределение эталона. Например, для эталонного предложения «Ученые открыли новую планету» и сгенерированного варианта «Астрономы обнаружили неизвестный объект» MoverScore оценит

семантическую близость между словами «учеными» и «астрономами», а также «планетой» и «объектом», несмотря на отсутствие точных лексических совпадений.

- Perplexity

Perplexity (Перплексия) отражает степень «удивления» языковой модели при генерации текста: чем ниже значение, тем лучше модель предсказывает последовательности

- Coherence Score

Coherence Score анализирует логическую связность текста с помощью методов, таких как тематическое моделирование (например, Latent Dirichlet Allocation, LDA) или построение графов связности, где узлы представляют ключевые слова, а рёбра — смысловые связи между предложениями. Пример текста с низким Coherence Score: «Кот сидит на ковре. Он любит молоко. Завтра будет дождь». В данном случае разрыв темы («дождь» не связан с «котом») снижает оценку связности

Сводная таблица характеристик метрик представлена в табл. 1.

ТАБЛИЦА I.

Метрика	**Основа**	**Учет семантики**	**Вычислительная сложность**
BLEU	n-граммы	Нет	Низкая
ROGUE-L	LCS	Нет	Средняя
BERTScore	Эмбеддинги BERT	Да	Высокая
MoverScore	EMD + Эмбеддинги	Да	Очень высокая
Preplexity	Вероятности модели	Нет	Низкая

B. Описание проведенного эксперимента

Для практической оценки метрик использовался датасет CNN/DailyMail 3.0.0 — стандартный бенчмарк для задач суммаризации текста, содержащий 300 тысяч новостных статей с аннотациями, написанными человеком.

Датасет включает статьи CNN и Daily Mail (2010–2015 гг.), где каждая запись содержит:

- article — исходный текст (500–1000 слов);
- highlights — краткие аннотации (3–5 предложений), созданные людьми.

Его преимущества — репрезентативность, стандартизация и наличие эталонов. Однако он смещён в сторону новостного стиля, что ограничивает применимость результатов к нишевым доменам, таким как медицина или программирование.

На тестовой выборке из 100 статей были сгенерированы суммаризации моделями BART (специализированная для суммаризации) и GPT-4 (общего назначения).

Результаты оценки представлены в табл. 2

C. Сравнительный анализ метрик

ТАБЛИЦА II.

Метрика	**BART**	**GPT-4**
BLEU	0.247	0.182
ROGUE-L	0.391	0.297
BERTScore	0.854	0.792

- BLEU: BART показала значение 0.247, GPT-4 — 0.182. Низкие абсолютные значения связаны с ориентацией метрики на точные лексические совпадения. Например, фразы «экономический рост составил 5%» и «ВВП увеличился на 5%» получат низкий BLEU, но высокий BERTScore.
- ROUGE-L: BART — 0.391, GPT-4 — 0.297. Метрика, учитывающая длинные совпадения, лучше отражает полноту, но игнорирует семантику.
- BERTScore: BART — 0.854, GPT-4 — 0.792. Высокие значения BART подтверждают её способность сохранять смысловую адекватность, даже при различиях в формулировках.

Специализация модели играет критическую роль: BART, обученная на задачах суммаризации, превзошла GPT-4 по всем метрикам, особенно по BERTScore (+7.8 %). Это связано с её архитектурой, оптимизированной для выделения ключевых тезисов.

Ограничения BLEU и ROUGE проявляются в их ориентации на лексические совпадения, что делает их менее релевантными для задач, требующих семантического анализа. Например, на разговорных данных (твиты, диалоги) эти метрики могут давать менее точные результаты из-за вариативности формулировок.

Для инженерных задач, где важна скорость, оптимальна комбинация BLEU и ROUGE. Для анализа смысловой корректности предпочтительны BERTScore или MoverScore. В исследовательских целях рекомендуется дополнять автоматические метрики экспертной оценкой, например, проверкой связности текста по шкале от 1 до 5.

Автоматические метрики эффективны для сравнения моделей в рамках одного домена, но их интерпретация требует учёта трёх факторов:

- Типа задачи (суммаризация, диалог, перевод).
- Специализации модели (BART для суммаризации vs. GPT-4 для генерации).
- Семантической глубины (формальные тексты vs. креативные формулировки).

Метрики, ориентированные на разные аспекты качества, демонстрируют неоднозначные результаты. Традиционные подходы, такие как BLEU и ROUGE, фокусируются на поверхностных совпадениях, что приводит к низким абсолютным значениям (0.18–0.39) в задачах, где важна семантика. Например, фразы «экономический рост» и «увеличение ВВП» получают близкие значения BERTScore (0.85), но нулевой BLEU из-за различий в формулировках. Это подчеркивает необходимость комбинированного использования

метрик: BLEU и ROUGE — для быстрой оценки, BERTScore — для анализа смысловой корректности.

Практические рекомендации основаны на учете доменной специфики и целей проекта. Для инженерных задач, таких как A/B-тестирование или мониторинг моделей, оптимальны BLEU и ROUGE, обеспечивающие скорость и простоту интерпретации. Однако в сценариях, где критична семантическая точность — например, при генерации медицинских заключений или юридических документов — предпочтение следует отдавать BERTScore или MoverScore. Эти метрики, несмотря на высокую вычислительную сложность, позволяют оценить контекстуальную адекватность текста, что особенно важно для нишевых доменов.

Адаптация метрик к конкретным данным требует проведения пилотной оценки. Например, при работе с медицинскими текстами необходимо проверить, как BERTScore реагирует на синонимы профессиональных терминов, и при необходимости скорректировать веса метрик. Для юридических документов рекомендуется разрабатывать эталонные аннотации, отражающие ключевые правовые формулировки, что повысит точность оценки.

Примеры применения иллюстрируют, как выбор метрик влияет на итоговый результат. В кейсе суммаризации медицинских исследований основным критерием становится ROUGE-L, оценивающий полноту ключевых терминов, а BERTScore используется для проверки контекстуальной корректности. В диалоговых системах, где важна естественность ответов, акцент смещается на Coherence Score и экспертные оценки, дополняющие автоматические методы.

Направления для будущих исследований включают разработку гибридных метрик, сочетающих преимущества формальных и семантических подходов.

Например, комбинация BERTScore с доменно-специфичными правилами (проверка корректности ссылок в академических текстах) могла бы повысить надежность оценки. Ещё одним перспективным направлением является динамическая калибровка метрик, позволяющая автоматически подбирать веса в зависимости от типа данных — новостей, диалогов или технической документации.

Сравнительный анализ подтвердил, что универсального подхода к оценке генерации не существует. Эффективность метрик зависит от контекста: BART демонстрирует превосходство в семантике, но для быстрого прототипирования может быть достаточно GPT-4 с ROUGE. Ключевая рекомендация — разработка гибких pipelines, объединяющих автоматические метрики, экспертизу и доменные знания. Это позволит обеспечить комплексную оценку качества генерации в реальных сценариях, от медицинских исследований до диалоговых систем.

II. Заключение

Развитие больших языковых моделей (LLM) кардинально изменило подходы к генерации текста, сделав актуальным вопрос объективной оценки качества их выводов. Проведенный сравнительный анализ автоматических метрик, таких как BLEU, ROUGE-L и BERTScore, на примере моделей BART и GPT-4, позволил выявить ключевые закономерности, сформулировать практические рекомендации и обозначить перспективные направления для дальнейших исследований.

Исследование подтвердило, что специализация моделей играет решающую роль в качестве генерации. Модель BART, оптимизированная для суммаризации, продемонстрировала значительное превосходство над GPT-4 общего назначения, особенно в семантических метриках. Например, разрыв в BERTScore составил 7.8% (0.854 против 0.792), что подчеркивает её способность сохранять смысловую адекватность даже при отсутствии точных лексических совпадений. Это связано с архитектурными особенностями BART, такими как механизм внимания, который эффективно выделяет ключевые тезисы из исходного текста.

Традиционные метрики, такие как BLEU и ROUGE, несмотря на их простоту и скорость вычисления, показали ограниченную применимость в задачах, требующих учета семантики. Их низкие абсолютные значения (0.18–0.39) объясняются ориентацией на поверхностные совпадения, что делает их менее релевантными для оценки перефразирования или контекстуальной связности. Например, фразы «экономический рост» и «увеличение ВВП» получили близкие значения BERTScore (0.85), но нулевой BLEU, что иллюстрирует принципиальное различие в подходах к оценке.

Практическая значимость исследования заключается в разработке рекомендаций по выбору метрик в зависимости от целей проекта. Для инженерных задач, таких как A/B-тестирование или мониторинг моделей в реальном времени, оптимальны BLEU и ROUGE, обеспечивающие скорость и простоту интерпретации. Однако в сценариях, где критична семантическая точность — например, при генерации медицинских заключений, юридических документов или диалоговых ответов — предпочтение следует отдавать BERTScore или MoverScore. Эти метрики, несмотря на высокую вычислительную сложность, позволяют оценить контекстуальную адекватность текста, что особенно важно для нишевых доменов.

Адаптация метрик к доменным данным требует проведения пилотной оценки и калибровки. Например, при работе с медицинскими текстами необходимо проверить, как BERTScore реагирует на синонимы профессиональных терминов, а для юридических документов — разработать эталонные аннотации, отражающие ключевые правовые формулировки. Внедрение таких подходов повысит точность оценки и снизит риски смысловых искажений.

Ограничения исследования связаны с выбором датасета CNN/DailyMail, который ориентирован на новостные статьи и может не отражать специфику других доменов. Кроме того, автоматические метрики, даже самые продвинутые, не заменяют экспертной оценки, особенно в задачах, требующих творческого подхода или глубокого понимания контекста.

Перспективные направления для будущих исследований включают разработку гибридных метрик, сочетающих преимущества формальных и семантических подходов. Например, интеграция BERTScore с доменно-специфичными правилами (проверка корректности цитирований в академических текстах) могла бы повысить надежность оценки. Динамическая калибровка метрик, позволяющая автоматически подбирать веса в зависимости от типа данных (новости, диалоги, техническая документация), также представляет значительный интерес. Отдельным направлением является улучшение интерпретируемости результатов — например, визуализация вклада отдельных слов в итоговый балл BERTScore, что упростит анализ ошибок генерации.

В заключение, сравнительный анализ подтвердил, что универсального подхода к оценке качества генерации не существует. Эффективность метрик зависит от контекста: BART демонстрирует превосходство в семантике, но для быстрого прототипирования может быть достаточно GPT-4 с ROUGE. Ключевая рекомендация — разработка гибких решений, объединяющих автоматические метрики, экспертизу и доменные знания. Такой подход обеспечит комплексную оценку качества генерации в реальных сценариях, от медицинских исследований до диалоговых систем, способствуя созданию более надежных и эффективных языковых моделей.

Несмотря на прогресс в разработке автоматических метрик, таких как BERTScore и MoverScore, экспертная оценка человеком остается незаменимым инструментом для анализа качества генерации текста. Автоматические методы, даже самые продвинутые, не способны полноценно оценить контекстуальную уместность, эмоциональную окраску или креативность текста. Например, в задачах генерации поэзии, маркетинговых слоганов или диалоговых ответов, где важны нюансы стиля и культурные отсылки, только человек может определить, насколько результат соответствует ожиданиям. Это делает экспертизу критически важной для валидации моделей в реальных сценариях.

Экспертная оценка особенно актуальна в нишевых доменах, где автоматические метрики сталкиваются с ограничениями. Например, в медицинских текстах некорректная интерпретация терминов может привести к фатальным ошибкам, а в юридических документах — к правовым рискам. Только специалисты, обладающие глубокими знаниями в предметной области, способны выявить смысловые неточности или противоречия, которые остаются незамеченными для алгоритмов. Кроме того, в креативных индустриях (литература, реклама) человеческая оценка становится ключевым критерием успеха, так как креативность и оригинальность трудно формализовать математически.

Одной из проблем экспертной оценки является субъективность, которая может искажать результаты. Для минимизации этого эффекта необходима разработка унифицированных протоколов оценки, включающих четкие критерии (например, шкалы для оценки связности, релевантности, грамматической корректности) и обучение экспертов. Например, в академических исследованиях часто используется метод слепой оценки, когда несколько независимых экспертов анализируют тексты без знания источника их генерации.

Перспективным направлением является гибридизация методов, где автоматические метрики и экспертные оценки дополняют друг друга. Например, BERTScore может использоваться для первоначальной фильтрации явно некорректных текстов, после чего эксперты фокусируются на анализе сложных случаев. Обратная связь от экспертов, в свою очередь, может применяться для тонкой настройки автоматических метрик, улучшая их чувствительность к доменно-специфичным нюансам. Исследования в этой области могли бы включать разработку алгоритмов активного обучения, где модель запрашивает экспертные оценки только для тех данных, где автоматические метрики демонстрируют низкую уверенность.

Для снижения нагрузки на экспертов и повышения эффективности их работы необходимы инструменты автоматизации рутинных задач. Например, платформы для аннотирования текстов с функциями подсветки спорных фрагментов, автоматической проверки грамматики или поиска противоречий. Полуавтоматические системы, где ИИ предлагает предварительные оценки, а эксперт их корректирует, могли бы ускорить процесс и снизить затраты. Ещё одним направлением является использование краудсорсинговых платформ для масштабирования экспертной оценки, однако здесь требуется решение проблем согласованности и контроля качества данных.

Список литературы

[1] Браун Т.Б., Манн Б., Райдэр Н., Саббих М., Каплан Дж., Дхаривал П., Нилакант А., Шям П., Састри Г., Аскелл А., Агарвал С., Херберт–Восс А., Крёгер Г., Хенайхан Т., Чайлд Р., Рамеш А., Зиглер Д., Ву Д., Винтер К., Хессе Ч., Чен М., Сиглер Э., Литвин М., Грей С., Чесс Б., Кларк Дж., Бернер К., МакКэндлиш С., Радфорд А., Сатскивер И., Амодеи Д. Language Models are Few-Shot Learners // ArXiv preprint arXiv:2005.14165, 2020. 40 с.

[2] Радфорд А., Нарасимхан К., Салиманс Т., Сатскивер И. Improving Language Understanding by Generative Pre-Training // OpenAI, 2018. 35 с.

[3] Лин Ч.-Й. ROUGE: Пакет для автоматической оценки качества аннотаций // Труды семинара «Text Summarization Branches Out», 2004. 25 с.

[4] Папинени К., Роукос С., Уорд Т., Чжу В.Ж. BLEU: Метод автоматической оценки качества машинного перевода // Труды 40-го ежегодного заседания Ассоциации вычислительной лингвистики, 2002. 30 с.

[5] Ретер Э., Бельц А. Компьютерная оценка генерации естественного языка: причины и методы // Journal of Artificial Intelligence Research, 2018. Т. 61, вып. 1. С. 1–50.

[6] Чжан Т., Кишор В., У Ф., Вайнбергер К.К., Артци Й. BERTScore: Оценка качества генерации текста с использованием BERT // International Conference on Learning Representations (ICLR), 2020. 28 с.

[7] Селлам Т., Дас Р., Го Дж., Парих А., Белинков Й., Хои С.К.Х. BLEURT: Обучение надежных метрик для оценки генерации текста // ArXiv preprint arXiv:2004.04619, 2020. 32 с.

[8] Гатт А., Крахмер Э. Обзор современного состояния генерации естественного языка: основные задачи, приложения и оценка // Journal of Artificial Intelligence Research, 2018. Т. 61, вып. 1. С. 1–40.

[9] Мэйхью Л., Лестер Б., Зеттлемойер Л. Many Hands Make Light Work: Открытая оценка генерации текста // Труды конференции EMNLP, 2020. 26 с.