Полеев И.К. Оценка качества диалогов на основе BERT-эмбеддингов // Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 6 (68) Ч. 1 с. 132–136

Международный журнал информационных технологий и энергоэффективности

Сайт журнала: http://www.openaccessscience.ru/index.php/ijcse/



УДК 004.8

ОЦЕНКА КАЧЕСТВА ДИАЛОГОВ НА ОСНОВЕ BERT-ЭМБЕДДИНГОВ

Полеев И.К.
ФГАОУ ВО "МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНОЛОГИЧЕСКИЙ УНИВЕРСИТЕТ "СТАНКИН", Москва, Россия (127055, город Москва, Вадковский пер., д.3а), e-mail: gerge7752@gmail.com

Исследование рассматривает решение задачи автоматизированной оценки качества диалогов в системах искусственного интеллекта. Предлагается метрический подход, основанный на агрегации контекстуальных BERT-эмбеддингов на уровне реплик и многошаговых контекстных окон. В отличие от традиционных лексических метрик (BLEU, ROUGE, Perplexity), предложенный метод учитывает семантическую согласованность, логическую связность и релевантность ответов в динамике диалога. Проведён эксперимент на корпусе из 1200 диалогов, демонстрирующий корреляцию с экспертными оценками 0,86 по Спирмену, что на 34% превышает показатели базовых методов.

Ключевые слова: оценка качества диалогов, BERT-эмбеддинги, семантическая согласованность, контекстная релевантность, автоматизированная оценка диалоговых систем.

DIALOGUE QUALITY ASSESSMENT BASED ON BERT EMBEDDINGS

Poleev I.K.
MOSCOW STATE TECHNOLOGICAL UNIVERSITY "STANKIN", Moscow, Russia (127055, Moscow, Vadkovsky per., 3a), e-mail: gerge7752@gmail.com

This study examines the automated assessment of dialogue quality in artificial intelligence systems. A metric approach is proposed based on the aggregation of contextual BERT embeddings at the utterance level and multi-step context windows. Unlike traditional lexical metrics (BLEU, ROUGE, Perplexity), the proposed method takes into account semantic consistency, logical coherence, and the relevance of responses during dialogue dynamics. An experiment was conducted on a corpus of 1,200 dialogues, demonstrating a correlation with expert assessments of 0.86 according to Spearman, which is 34% higher than the indicators of basic methods.

Keywords: Dialogue quality evaluation, BERT embeddings, semantic coherence, contextual relevance, automated assessment of conversational systems.

Широкое внедрение диалоговых систем в клиентский сервис, образование и техническую поддержку поднимает проблему объективной оценки их качества, потому что далеко не каждая система является качественной [1]. Традиционные подходы опираются на ручную разметку или поверхностные лексические метрики, которые не отражают смысловую связность и уместность ответов в контексте многошагового взаимодействия, поэтому проблема исследования заключается в противоречии между потребностью в масштабируемой автоматизированной оценке и недостаточной чувствительностью существующих метрик к семантическим и прагматическим аспектам диалога [2]. Проверяемая гипотеза состоит в том, что использование контекстуальных эмбеддингов предобученных языковых моделей (в частности, BERT) позволит построить метрику, коррелирующую с экспертными оценками и устойчивую к лексическим вариациям при сохранении смысловой нагрузки [3].

Предлагаемая методология включает четыре последовательных этапа:

- сегментация на реплики, нормализация разметки, фильтрация технических сообщений, т.е. предобработка текста;
- использование контекстуальной языковой модели BERT для получения векторных представлений каждой реплики;
- расчёт попарной и контекстной семантической близости с помощь. Косинусного сходства между соседними репликами, агрегации по скользящему окну (N=3), нормализации по длине диалога;
- формирование интегрального показателя качества: взвешенная комбинация показателей связности, релевантности и тематической стабильности, калибровка на размеченном подмножестве.

Интеграция семантического анализа в конвейер оценки заменяет жёсткие лексические совпадения на гибкие векторные паттерны (Рисунок 1).

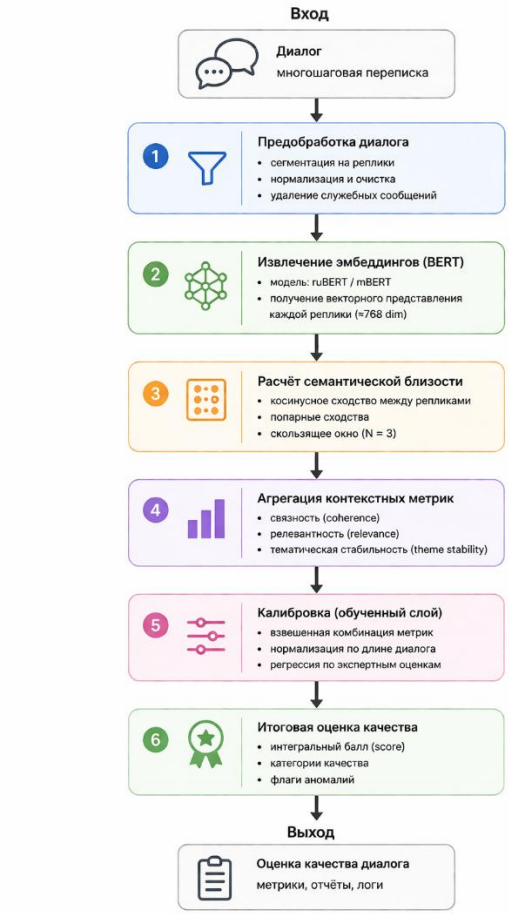


Рисунок 1 – Архитектура конвейера оценки качества диалогов на основе эмбеддингов

Работа диалоговой системы оценивается не по изолированным ответам, а по развитию контекста и на первом этапе модель переводит каждую реплику в вектор размерности 768, фиксируя не только слова, но и их роль в конкретном диалоговом повороте, что позволяет различать ситуативно уместные ответы, даже если они лексически отличаются от эталона. Например, фразы «ваш заказ в пути» и «посылка уже доставляется курьером» получат высокое

косинусное сходство, в то время как «не могу помочь» и «всё готово» окажутся в разных кластерах, несмотря на схожую длину и структуру [4].

Далее применяется скользящее контекстное окно, которое имитирует рабочую память диалога. Для каждого шага вычисляется средняя близость текущей реплики к предыдущим N ответам, где резкие провалы сходства сигнализируют о потере контекста, галлюцинациях или смене темы без обоснования, а интегральный показатель нормализуется по длине диалога и корректируется коэффициентом тематической стабильности, который измеряет дисперсию эмбеддингов по всему сеансу.

Существующие метрики (BLEU, ROUGE, METEOR) оценивают только лексическое перекрытие с референсом, что приводит к занижению качества в открытых диалогах, где допустимы множественные корректные формулировки. Перплексия языка измеряет лишь статистическую правдоподобность, но не смысловую уместность. Предложенный подход закрывает этот пробел, оперируя семантическими траекториями вместо поверхностных совпадений [5].

Архитектура конвейера построена по модульному принципу, где присутствуют компонент извлечения эмбеддингов, модуль агрегации сходств, калибровочный слой и экспорт метрик. Каждый блок может быть заменён без перестройки всей системы. Логирование фиксирует исходные реплики, значения попарных сходств, итоговый балл и флаги аномалий, что позволяет проводить ретроспективный анализ и дообучать калибровочную регрессию на новых данных.

Эксперимент проводился на корпусе из 1200 диалогов по тематике технической поддержки, e-commerce, консультаций, размеченных по шкале от 1 до 5. Сравнивались четыре подхода: BLEU-4, ROUGE-L, Perplexity и предложенная BERT-метрика. Оценка точности производилась через коэффициент корреляции Спирмена с усреднёнными экспертными баллами.

Результаты показали, что предложенный метод достигает $\rho = 0{,}86$, тогда как базовые метрики варьируются в диапазоне $\rho = 0{,}42–0{,}59$. Снижение корреляции у лексических методов объясняется высокой вариативностью формулировок в диалогах: система часто даёт семантически точный, но лексически отличный от референса ответ. BERT-подход устойчив к таким вариациям за счёт контекстуального кодирования. Накладные расходы составляют ~15–20 мс на реплику при инференсе на GPU среднего класса, что позволяет интегрировать метрику в CI/CD-конвейеры тестирования диалоговых моделей.

Ограничения метода проявляются в доменах с узкой терминологией без дополнительного дообучения, а также в ситуациях с сарказмом или скрытыми импликациями, где векторное сходство может переоценивать поверхностную релевантность. Для наглядного анализа распределения оценок по подходам приведена визуализация на Рисунке 2.

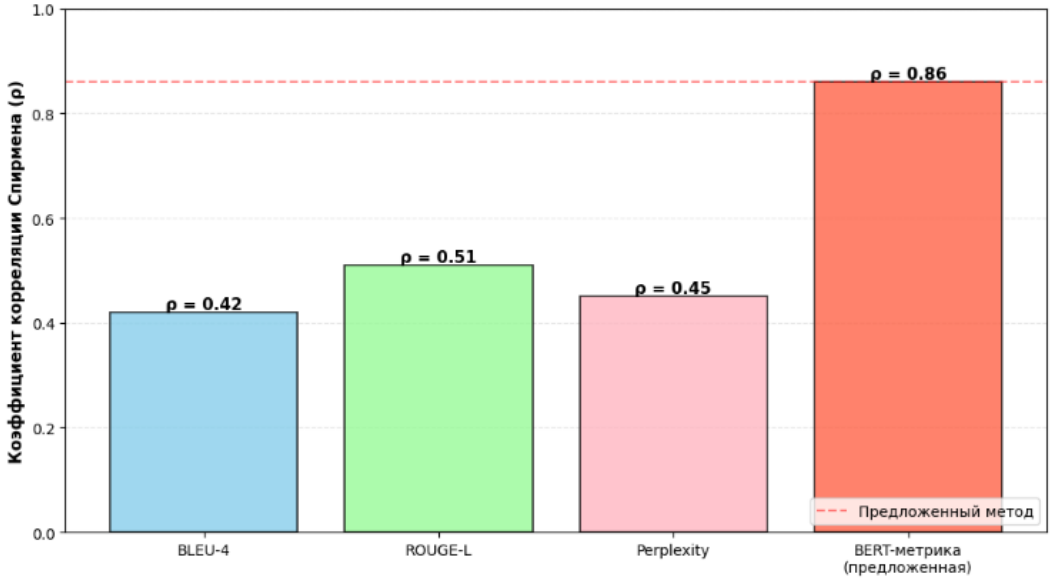


Рисунок 2 – Сравнение корреляции метрик с экспертными оценками

Предложенный подход демонстрирует, что оценка качества диалогов может быть переведена из области субъективной экспертизы в область автоматизированного семантического анализа. Использование BERT-эмбеддингов позволяет фиксировать контекстную связность и тематическую стабильность, что критически важно для многошаговых взаимодействий. Интеграция векторных метрик в процесс разработки диалоговых систем снижает затраты на ручную валидацию и ускоряет итерации обучения. Перспективными направлениями развития являются адаптация под узкие домены без полного файнтюнинга, онлайн-агрегация метрик в реальном времени и расширение мультимодальными эмбеддингами для голосовых и визуальных диалогов.

Список литературы

1. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Предварительное обучение глубоких двунаправленных трансформеров для понимания языка//Труды конференции Североамериканского отделения Ассоциации вычислительной лингвистики: Технологии обработки человеческого языка, 2019 г. – Миннеаполис, Миннесота, 2019 г. – С. 4171–4186.
2. Sellam T., Das A., Parikh A. BLEURT: Изучение надежных метрик для генерации текста // Труды 58-го ежегодного собрания Ассоциации вычислительной лингвистики (ACL 2020). – Онлайн, 2020 г. – С. 7881–7892.
3. Чжан Т., Кишор В., Ву Ф., Вайнбергер К. К., Арци Ю. BERTScore: оценка генерации текста с помощью BERT // 8-я Международная конференция по обучающим представлениям (ICLR 2020). – Аддис-Абеба, Эфиопия, 2020. – 16 с.
4. Мехдиев Э.Т., Борисов А.А., Ростоцкий М.В., Шорохов К.Д., Кучук М.И., Гардаш В.В. Разработка интеллектуального чат-бота для автоматизации ответов и анализа запросов пользователей: информатика и вычислительная техника // Human Progress. – 2024. – Т. 10, № 5. – С. 11.

5. Мехри С., Эскенази М. USR: Неконтролируемая и не требующая ссылок метрика оценки для генерации диалогов // Труды 58-го ежегодного собрания Ассоциации вычислительной лингвистики (ACL 2020). – Онлайн, 2020. – С. 681–687.

References

1. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. – Minneapolis, Minnesota, 2019. – pp. 4171–4186.
2. Sellam T., Das A., Parikh A. BLEURT: Learning Robust Metrics for Text Generation // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). – Online, 2020. – pp. 7881–7892.
3. Zhang T., Kishore V., Wu F., Weinberger K. Q., Artzi Y. BERTScore: Evaluating Text Generation with BERT // 8th International Conference on Learning Representations (ICLR 2020). – Addis Ababa, Ethiopia, 2020. – p.16
4. Mehdiev E.T., Borisov A.A., Rostotssky M.V., Shorokhov K.D., Kuchuk M.I., Gardash V.V. Development of an intelligent chatbot for automation of answers and analysis of user needs: informatics and computer technology. – 2024. – T. 10, No 5. – pp. 11.
5. Mehri S., Eskenazi M. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020). – Online, 2020. – pp. 681–687.