математических наук, доцент кафедры вычислительных методов механики деформируемого тела, e-mail: n.gasratova@spbu.ru
Евлахова Наталья Алексеевна, «Клиника эстетической хирургии «Абриелль»» г. Санкт-Петербург, Российская Федерация, пластический хирург, e-mail: dr.evlakhova@abriell.ru
Осаула Анастасия Николаевна, «Клиника эстетической хирургии «Абриелль»» г. Санкт-Петербург, Российская Федерация, пластический хирург, e-mail: dr.osaula@abriell.ru
Грецкова Евгения Евгеньевна, «Клиника эстетической хирургии «Абриелль»» г. Санкт-Петербург, Российская Федерация, пластический хирург, e-mail: dr.gretskova@abriell.ru

УДК 004.81

ОЦЕНКА КАЧЕСТВА ПЕРЕВОДА ДАТАСЕТА ДЛЯ ОЦЕНКИ ЭМПАТИИ: ГИПОТЕЗЫ И ЭКСПЕРИМЕНТЫ

Валиева Н.Л. [1]***, Буянов И.О.*** [2]***, Гончарова А.Б.*** [1]

[1]***«Санкт-Петербургский государственный университет», г. Санкт-Петербург***
[2]***Федеральный исследовательский центр «Информатика и управление» РАН, г. Москва***

Работа посвящена исследованию методов автоматической оценки качества перевода на русский язык англоязычного набора данных EPITOME [1], который состоит из текстов пользователей Reddit, размеченных по типам эмпатии и соответствующим носителям эмпатичных высказываний. Перевод выполнялся с помощью больших языковых моделей (LLM) в несколько этапов, а итоговый набор данных использовался для обучения моделей классификации эмпатии. Однако по результатам работы вопрос оценки качества перевода остался открытым. Даже при многоступенчатом пайплайне с участием нескольких моделей (YandexGPTPro, Qwen-2.5-72b, GPT-4o) качество отдельных примеров могло значительно меняться.

Для автоматического выявления неудачных переводов была проверена гипотеза о связи между перплексией текста, рассчитанной на русскоязычной языковой модели, и субъективной оценкой качества перевода. Для расчета перплаксиии спользовалась модель для русского языка, разработанная командой SberDevices [3].Были рассчитаны значения логарифмированной перплексии для всего переведенного набора данных. Наблюдалось распределение, с положительной асимметрией, положительный хвост интерпретировался как зона потенциально плохих

переводов. Для исключения зависимости перплексии от длины текста, тексты разделили на десять частей и вычисляли коэффициент внутри каждой части. Коэффициент для переводов, попавший в четвертый квартиль или правее него, относились к категории «плохих».

Три профессиональных переводчика с филологическим образованием и опытом работы более 2 лет оценивали качество переведенных текстов по пятибалльной шкале, опираясь на следующие критерии: грамматическая корректность, точность передачи смысла, естественность и связность речи, соответствие стилистике оригинала. Средние оценки сравнивались с помощью U-критерия Манна-Уитни. Полученное значение $p=0.08$ свидетельствует о возможной, но статистически неубедительной связи между перплексией и качеством перевода.

Проверена практическая гипотеза, что перевод лучше при использовании более крупной модели. Для этого были отобраны 200 «плохих» примеров, переведенных заново с помощью GPT-4o. Профессиональные переводчики сравнивали пары (старый/новый) и отмечали лучший вариант. Результаты анализировались биномиальным тестом, который показал высокую значимость ($p=0.001$), что подтверждает: использование более мощной модели действительно улучшает качество перевода.

Однако дальнейший анализ выявил низкую согласованность между переводчиками (коэффициент Криппендорфа −0.09), что указывает на высокую субъективность человеческих суждений и ставит под сомнение достоверность полученных ранее результатов.

Проверялась возможность использования LLM-as-a-Judge для шести моделей (GPT-4o, Claude-Sonnet-4, Gemini-2.5-Flash, Qwen-3-235B, DeepSeek-Chat-v3, Llama-4-Maverick) в двух режимах: численная оценка по пятибалльной шкале, бинарная оценка(«плохой»/«хороший»).Связь между оценками LLM и людьми проверялась χ^2-тестом. Результаты показали отсутствие значимых ассоциаций у большинства моделей.

Вывод работы: без надёжного эталона человеческих оценок невозможно валидировать автоматические метрики качества перевода. Предлагается поиск более устойчивых подходов к оценке, включая комбинирование статистических метрик и оценок нескольких LLM.

Литература

1. Sharma A., Miner A., Atkins D., Althoff T.A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). - Online: Association for Computational Linguistics, 2020. - С. 5263-5276.

2. Валиева Н.Л., Буянов И.О., Гончарова А.Б. Перевод на русский язык датасета для оценки эмпатии // Процессы управления и устойчивость. - 2025. - №S2-1. - С. 229-230.
3. Модель ai-forever/rugpt3small_based_on_gpt2 [Электронный ресурс] Режим доступа: https://huggingface.co/ai-forever/rugpt3small_based_on_gpt2 (Дата: 01.11.2025).

Валиева Нафиса Ленисовна, Санкт-Петербургский государственный университет г. Санкт-Петербург, Российская Федерация, бакалавр 3 курса, кафедра Теории систем управления электрофизической аппаратурой, e-mail: nafisalenva@gmail.com
Буянов Игорь Олегович, Федеральный Исследовательский центр "Информатика и управление" Российской Академии Наук, Москва, аспирант, e-mail: buyanov.igor.o@yandex.ru
Гончарова Анастасия Борисовна, Санкт-Петербургский государственный университет, г. Санкт-Петербург, Российская Федерация, кандидат физико-математических наук, доцент кафедры Теории систем управления электрофизической аппаратурой, e-mail: a.goncharova@spbu.ru

УДК 625.7

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ СИСТЕМЫ УПРАВЛЕНИЯ ПРОЦЕССОМ РАЗОГРЕВА БИТУМА НА ОСНОВЕ НЕЧЕТКОЙ ЛОГИКИ

Волков И.Н., Бурковский В.Л.

«Воронежский государственный технический университет», г. Воронеж, Россия

Современное дорожное строительство предъявляет высокие требования к качеству асфальтобетонных покрытий, которое напрямую зависит от свойств асфальтобетонной смеси. Ключевым параметром, определяющим эти свойства, является битум. Нарушение температурного режима приводит к ухудшению характеристик битума, что, в свою очередь, вызывает недостаточное уплотнение покрытия, образование трещин и преждевременное разрушение дорожного полотна. Оптимальный диапазон температуры выходящего битума для производства качественной смеси находится в пределах 150-170°С[1]. Поддержание этого параметра непростая задача, так как технологический процесс разогрева подвержен влиянию множества факторов неопределенности. Традиционные системы автоматического регулирования не всегда эффективно справляются с объектами, обладающими нелинейностью и неопределенностью, каким