**УДК 005.94**

**Ф.В. Краснов, кандидат технических наук, fedor.krasnov@vseinstrumenti.ru, ООО «Ви.Тех», г.Москва**

# Легковесные алгоритмы коррекции раскладки ENG→RU в условиях высокой нагрузки

**Аннотация:** Современные поисковые системы электронной коммерции сталкиваются с трудностями обработки запросов, содержащих короткие технические токены, особенно в сегменте «сделай сам» (DIY). Одной из ключевых проблем является ошибка раскладки клавиатуры (ENG→RU), приводящая к искажению и нераспознаванию запросов, например, «ыыВ 1ЕИ» вместо «SSD 1TB». Подобные искажения снижают точность поиска и конверсию. В работе предложены легковесные алгоритмы коррекции раскладки, оптимизированные для высоконагруженных систем с ограниченными вычислительными ресурсами. Эксперименты на доменных датасетах демонстрируют рост точности поиска на 25–30% при времени отклика менее 10 мс на запрос. Полученные результаты подтверждают применимость подхода в промышленных IR-системах и облачных платформах.

**Ключевые слова:** информационный поиск, электронная коммерция, раскладка клавиатуры, транслитерация, высокая нагрузка, ограниченные ресурсы, DIY.

## 1. Введение

Платформы электронной коммерции, особенно в сегменте «сделай сам», часто обрабатывают короткие запросы с техническими обозначениями — «bolt M8x50», «drill 18V» и т.п. Ошибки раскладки клавиатуры (ENG→RU) приводят к тому, что латинские символы интерпретируются как кириллические, делая запросы нераспознаваемыми для поискового индекса. Это снижает

релевантность и конверсию до 40–50%, согласно аналитике крупных маркетплейсов (Wildberries, Ozon).

Цель исследования — разработка и оценка вычислительно эффективных алгоритмов коррекции раскладки для высоконагруженных IR-систем. В отличие от орфографических методов, ориентированных на естественный язык, подход нацелен на восстановление коротких технических токенов. Работа объединяет контекстное сопоставление, эвристические правила и лёгкие модели машинного обучения, формируя основу для масштабируемой обработки запросов в реальном времени.

## 2. Обзор литературы

Проблема обработки шумных запросов активно исследуется в контексте поисковых систем и машинного перевода. Ранние работы по статистической транслитерации [1] и коррекции имён собственных [2] улучшали поиск в многоязычных сценариях, но не учитывали визуальные ошибки клавиатурной раскладки. Многоязычные интерфейсы и модели ввода [3, 4] рассматривали транслитерацию для индийских языков, однако их применение к ENG→RU ограничено.

Для русского языка проводились исследования по коррекции орфографии [5, 8], устойчивые к шуму, но не адаптированные к коротким токенам и техническим обозначениям. Нейронные подходы [6, 7] обеспечивают высокую точность, но требуют значительных вычислительных ресурсов, что делает их непрактичными в условиях реального времени. Работа [11], ориентированная на моделирование произношения для улучшения орфографической коррекции, демонстрирует эффективность для английского языка, но её фонетическая направленность ограничивает применимость к визуальным ошибкам раскладки типа ENG→RU. В электронной коммерции подтверждено влияние корректировки запросов на рост конверсии [10], однако специфика коротких технических токенов в сегменте DIY остаётся недостаточно изученной.

Таким образом, существующие решения, включая Yandex Speller [8], не масштабируются под серверные сценарии с высокой нагрузкой. Расширение запросов через векторные представления слов [9] улучшает релевантность, но редко учитывает ограничения ресурсов.

Для коротких технических токенов коррекция орфографии малоприменима, так как транслитерированные ошибки не соответствуют лингвистическим паттернам [12]. Алгоритмы раскладочной коррекции, напротив, детерминированны и эффективно восстанавливают исходный токен на основе фиксированных таблиц сопоставления [13]. Это определяет актуальность легковесных решений, устойчивых к шуму и пригодных для промышленного внедрения.

## 3. Методология

Методология реализует многоуровневый подход к коррекции раскладки клавиатуры в пайплайнах информационного поиска. На первом этапе выполняется детекция ошибок — определение токенов, требующих исправления. Для этого используется компактная нейронная модель, обученная на доменных данных сегмента «сделай сам».
Средний словарь каталога насчитывает 1–3 млн токенов, а с учётом длинного хвоста низкочастотных комбинаций их число в запросах достигает 10 млн в сутки. Распределение отклоняется от закона Ципфа, что требует динамических фильтров и отказа от статичных словарей.

Ключевой элемент методологии — разграничение задач **обнаружения** и **исправления** транслитерации.

- Обнаружение фокусируется на токенах, возникших из-за ошибочной раскладки ENG→RU, и использует паттерны несоответствия кириллических и латинских символов.

- Исправление выполняется по фиксированным маппингам клавиш, восстанавливая исходный смысл токена.

Контекст запроса используется для минимизации ложных срабатываний: например, «ЫЫВ» может обозначать «SSD», но также трактоваться как аббревиатура. Категориальные признаки и соседние токены позволяют уточнять интерпретацию.

Контекстное сопоставление задействует векторные представления и анализ n-грамм для оценки семантической согласованности с каталогом товаров. Особое значение контекстное сопоставление приобретает в сегменте «сделай сам», где короткие технические обозначения (например, M8x50) нередко совпадают по форме с транслитерированными ошибками (Ь8ч50). Для предотвращения ложноположительных срабатываний алгоритм оценивает их согласованность с внутренними данными каталога и релевантными товарными категориями. Пост-валидация подтверждает, что исправленный токен релевантен документам индекса.

Метод поддерживает частичную коррекцию: в запросах типа «купить LED лампу» корректируется лишь один токен. Это обеспечивает адаптивность в реальных пользовательских сценариях. Итеративный цикл — **детекция → исправление → валидация** — обеспечивает баланс точности и производительности при минимальных вычислительных затратах.

## 4. Эксперимент и результаты

### 4.1 Данные

Для проверки эффективности предложенных алгоритмов сформирован корпус из реальных и синтетических поисковых запросов сегмента DIY. Основу составили данные открытых каталогов (Wildberries, Ozon). Аугментация выполнялась по фиксированным маппингам ENG→RU («S»→«Ы», «D»→«В»), с варьированием количества искажённых токенов.

Итоговый корпус содержит около **10 млн** записей, сбалансированных по категориям («инструменты», «электроника» и др.). Распределение по числу ошибок демонстрирует выраженный «длинный хвост»: большинство запросов корректны, но редкие сильно искажённые примеры критически влияют на качество поиска.

*Таблица 1: Распределение классов ошибок*

| Количество ошибок в поисковом запросе | Доля в датасете |
|---|---|
| 0 | 0.2063 |
| 1 | 0.0312 |
| 2 | 0.0727 |
| 3 | 0.1173 |
| 4 | 0.1523 |
| 5 | 0.1674 |
| 6 | 0.1463 |
| 7 | 0.0844 |
| 8 | 0.0222 |

Таблица 1 демонстрирует распределение классов по количеству ошибок раскладки в поисковых запросах. Наибольшую долю составляют корректные или слабо искажённые запросы (0–2 ошибки), в то время как многократные ошибки (5 и более) встречаются значительно реже. Такая асимметрия отражает типичную структуру пользовательских данных и подчёркивает необходимость алгоритмов, устойчивых к редким, но критически значимым случаям искажений.

Корпус разделён на обучающую (70%), валидационную (15%) и тестовую (15%) выборки.

### 4.2 Экспериментальная постановка

Оценка проводилась в условиях, имитирующих реальные IR-системы: нагрузка до 1000 запросов/сек на облачных узлах с ограниченными ресурсами.
Сравнивались три подхода:

1. стандартная орфографическая коррекция (базовый уровень),
2. эвристический алгоритм раскладочной коррекции,
3. гибридная модель (эвристика + ML).

Метрики включали точность, полноту, F1-меру, а также пропускную способность и среднюю задержку отклика.

### 4.3 Результаты

Гибридная модель показала рост релевантности на **25–30%** по сравнению с базовым методом. Среднее время отклика — **8–9 мс** (95-й перцентиль < 17 мс), что соответствует требованиям SLA для коммерческих платформ. Пропускная способность достигала **115 запросов/сек** при нагрузке без сбоев.
Использование легковесной модели fastText [17] и оптимизированных структур данных в памяти обеспечило загрузку ЦПУ < 70%.

По сравнению с Yandex Speller производительность возросла на 350–400%, при этом точность исправления коротких токенов увеличилась с 0.76 до 0.97 (F1). Эксперимент подтвердил эффективность предложенного подхода для облачных и мобильных систем с ограниченными ресурсами.

## 5. Обсуждение

Результаты демонстрируют, что комбинированный подход обеспечивает высокий уровень детерминизма при минимальном потреблении ресурсов. Отказ от тяжёлых трансформерных архитектур (T5 [14], BART[16]) позволил избежать задержек >50 мс, типичных для нейросетевых моделей. Субсловная

токенизация SentencePiece [15] оказалась непригодной для коротких технических обозначений («M8x50», «LED5W»), нарушая атомарность токенов.

Таким образом, оптимальным решением для высоконагруженных систем является сочетание:

- фиксированных таблиц маппинга для ENG→RU,
- контекстного сопоставления на уровне n-грамм,
- и лёгких моделей fastText [17], интегрированных в компилируемое окружение Go.

Рост метрики F1 на отложенной выборке с 90% до 97% был достигнут за счёт оптимизации гиперпараметров всего конвейера подготовки данных и обучения модели.

Последующая квантизация позволила уменьшить размер модели примерно на 30% без снижения качества предсказаний.

Следует отметить, что аналогичные улучшения трудно воспроизвести в трансформерных архитектурах, где высокая параметрическая сложность и чувствительность к масштабированию существенно ограничивают эффективность подобных оптимизаций.

Подобная архитектура гарантирует воспроизводимость и устойчивость к флуктуациям трафика при одновременном соблюдении требований SLA < 20 мс.

## 6. Заключение

Работа представила вычислительно эффективный подход к коррекции раскладки клавиатуры ENG→RU в поисковых системах электронной коммерции. Метод объединяет эвристические правила и лёгкое машинное обучение для восстановления технических токенов в реальном времени.

Эксперименты показали увеличение точности поиска на 25–30% и сокращение времени отклика до < 10 мс. Подход доказал масштабируемость и применимость в системах с ограниченными ресурсами, обеспечивая устойчивую работу под высокой нагрузкой.

Перспективы включают расширение набора раскладок (RU→ENG), внедрение самообучения и исследование влияния коррекции на downstream-задачи — расширение запросов, ранжирование и персонализацию поиска.

**Список литературы**


1. Lee J. S., Choi K. S. English to Korean statistical transliteration for information retrieval //Computer Processing of Oriental Languages. – 1998. – Т. 12. – №. 1. – С. 17-37.
2. Pogrebnoi D., Funkner A., Kovalchuk S. RuMedSpellchecker: Correcting Spelling Errors for Natural Russian Language in Electronic Health Records Using Machine Learning Techniques //International Conference on Computational Science. – Cham : Springer Nature Switzerland, 2023. – С. 213-227.
3. Raj A. A. A. Multi-lingual Screen Reader and Processing of Font-data in Indian languages : дис. – MS Thesis at International Institute of Information Technology Hyderabad, India, 2008.
4. Prabhakar D. K., Pal S. Machine transliteration and transliterated text retrieval: a survey //Sādhanā. – 2018. – Т. 43. – №. 6. – С. 93.
5. Balabaeva K., Funkner A., Kovalchuk S. Automated spelling correction for clinical text mining in Russian //Digital Personalized Health and Medicine. – IOS press, 2020. – С. 43-47.
6. Chari A., Ounis I., MacAvaney S. Lost in Transliteration: Bridging the Script Gap in Neural IR //Proceedings of the 48th International ACM SIGIR

Conference on Research and Development in Information Retrieval. – 2025. – С. 2900-2905.

7. Bruch S. et al. Special Section on Efficiency in Neural Information Retrieval // ACM Transactions on Information Systems. – 2024. – Т. 42. – №. 5. – С. 1-4.
8. Rozovskaya A. Spelling correction for Russian: A comparative study of datasets and methods //Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). – 2021. – С. 1206-1216.
9. Mikolov T. et al. Efficient estimation of word representations in vector space // arXiv preprint arXiv:1301.3781. – 2013.
10. Sachdeva N., McAuley J. How useful are reviews for recommendation? a critical review and potential improvements //proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval. – 2020. – С. 1845-1848.
11. Toutanova K., Moore R. C. Pronunciation modeling for improved spelling correction //Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. – 2002. – С. 144-151.
12. Belinkov Y., Bisk Y. Synthetic and natural noise both break neural machine translation //arXiv preprint arXiv:1711.02173. – 2017.
13. Müller L. et al. Dictionary Attack with Transformed Russian Words using QWERTY Keyboard Layout. – 2024.
14. Xue L. et al. mT5: A massively multilingual pre-trained text-to-text transformer //arXiv preprint arXiv:2010.11934. – 2020.
15. Kudo T., Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing //arXiv preprint arXiv:1808.06226. – 2018.

16. Lewis M. et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension //arXiv preprint arXiv:1910.13461. – 2019.

17. Joulin A. et al. Bag of tricks for efficient text classification //arXiv preprint arXiv:1607.01759. – 2016.

F.V.Krasnov, Candidate of Technical Sciences, fedor.krasnov@vseinstrumenti.ru,
LLC «Vi.Tech», Moscow

## Lightweight Algorithms for ENG→RU Keyboard Layout Correction under High-Load Conditions

**Abstract:** Modern e-commerce search systems increasingly encounter challenges in processing user queries containing short, domain-specific technical tokens, particularly within the “do-it-yourself” (DIY) segment. A major source of retrieval errors arises from incorrect keyboard layouts (ENG→RU), which produce distorted or unrecognized queries—such as “ыыВ 1ЕИ” instead of “SSD 1TB.” These distortions disrupt lexical and semantic integrity, significantly degrading retrieval accuracy and user conversion rates.

This study presents a set of lightweight algorithms for ENG→RU keyboard layout correction, specifically optimized for deployment in high-load information retrieval (IR) systems operating under limited computational resources. The proposed methods integrate deterministic key-mapping with heuristic and contextual analysis to restore query integrity in real time. Experimental evaluations on domain-specific e-commerce datasets demonstrate a 25–30% improvement in retrieval accuracy and an average response latency below 10 ms per query. The findings confirm the effectiveness and scalability of the approach for industrial and cloud-based IR applications.

**Keywords:** information retrieval, e-commerce, keyboard layout correction, transliteration, high-load systems, limited computing resources, DIY.

**References:**

1. Lee J. S., Choi K. S. English to Korean statistical transliteration for information retrieval //Computer Processing of Oriental Languages. – 1998. – Т. 12. – №. 1. – С. 17-37.

2. Pogrebnoi D., Funkner A., Kovalchuk S. RuMedSpellchecker: Correcting Spelling Errors for Natural Russian Language in Electronic Health Records Using Machine Learning Techniques //International Conference on Computational Science. – Cham : Springer Nature Switzerland, 2023. – С. 213-227.
3. Raj A. A. A. Multi-lingual Screen Reader and Processing of Font-data in Indian languages : дис. – MS Thesis at International Institute of Information Technology Hyderabad, India, 2008.
4. Prabhakar D. K., Pal S. Machine transliteration and transliterated text retrieval: a survey //Sādhanā. – 2018. – Т. 43. – №. 6. – С. 93.
5. Balabaeva K., Funkner A., Kovalchuk S. Automated spelling correction for clinical text mining in Russian //Digital Personalized Health and Medicine. – IOS press, 2020. – С. 43-47.
6. Chari A., Ounis I., MacAvaney S. Lost in Transliteration: Bridging the Script Gap in Neural IR //Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. – 2025. – С. 2900-2905.
7. Bruch S. et al. Special Section on Efficiency in Neural Information Retrieval // ACM Transactions on Information Systems. – 2024. – Т. 42. – №. 5. – С. 1-4.
8. Rozovskaya A. Spelling correction for Russian: A comparative study of datasets and methods //Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021). – 2021. – С. 1206-1216.
9. Mikolov T. et al. Efficient estimation of word representations in vector space // arXiv preprint arXiv:1301.3781. – 2013.
10. Sachdeva N., McAuley J. How useful are reviews for recommendation? a critical review and potential improvements //proceedings of the 43rd

international ACM SIGIR conference on research and development in information retrieval. – 2020. – С. 1845-1848.

11. Toutanova K., Moore R. C. Pronunciation modeling for improved spelling correction //Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. – 2002. – С. 144-151.
12. Belinkov Y., Bisk Y. Synthetic and natural noise both break neural machine translation //arXiv preprint arXiv:1711.02173. – 2017.
13. Müller L. et al. Dictionary Attack with Transformed Russian Words using QWERTY Keyboard Layout. – 2024.
14. Xue L. et al. mT5: A massively multilingual pre-trained text-to-text transformer //arXiv preprint arXiv:2010.11934. – 2020.
15. Kudo T., Richardson J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing //arXiv preprint arXiv:1808.06226. – 2018.
16. Lewis M. et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension //arXiv preprint arXiv:1910.13461. – 2019.
17. Joulin A. et al. Bag of tricks for efficient text classification //arXiv preprint arXiv:1607.01759. – 2016.