

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ЭФФЕКТИВНОСТИ МЕТОДОВ ПОИСКА И ГЕНЕРАЦИИ ОТВЕТОВ В ДИАЛогоВЫХ СИСТЕМАХ

Гофуржонов Мухаммадали Расулжон угли

ассистент, кафедра Искусственного интеллекта,

Ташкентский университет информационных технологий.

Тел.: +998 (97) 726-66-26

<https://orcid.org/0009-0003-2014-8248>

Бурханова Айгуль Ильясвовна

магистр 1-го курса, факультет Магистратуры, специальность «Искусственный интеллект»,

Ташкентский университет информационных технологий.

Тел.: +998 (50) 882-15-33 aygulburkhanovaluck@gmail.com

<https://doi.org/10.5281/zenodo.17832907>

Аннотация. В работе проведено сравнительное исследование трех методов формирования ответов в диалоговых системах: TF-IDF, BM25 и генеративного подхода.

Результаты демонстрируют превосходство BM25 по показателю релевантности ответов (4.40/5.0) при оптимальном времени выполнения (0.156 мс), в то время как генеративный подход показал наивысшую скорость работы (0.010 мс) при снижении качества ответов.

Автоматические метрики BLEU и ROUGE показали ограниченную эффективность в оценке качества диалоговых систем [1, 2].

Ключевые слова: диалоговые системы, информационный поиск, BM25, TF-IDF, генеративные модели, оценка качества, релевантность, время ответа

Abstract. This study presents a comparative analysis of three response-generation methods in dialogue systems: TF-IDF, BM25, and a generative approach. The results demonstrate the superiority of BM25 in terms of response relevance (4.40/5.0) with optimal execution time (0.156 ms), while the generative approach achieved the highest processing speed (0.010 ms) at the cost of reduced answer quality.

The automatic BLEU and ROUGE metrics showed limited effectiveness in evaluating the quality of dialogue systems [1, 2].

Keywords: dialogue systems, information retrieval, BM25, TF-IDF, generative models, quality evaluation, relevance, response time.

Введение

Современные диалоговые системы требуют баланса между скоростью ответа и его качеством [3]. Актуальной задачей является выбор оптимального метода обработки пользовательских запросов, обеспечивающего как высокую релевантность ответов, так и минимальное время отклика.

В исследовании сравниваются три подхода: классические методы поиска (TF-IDF [4], BM25 [5]) и правило-базируемый генеративный метод для определения наиболее эффективного решения.

Методы

Исследование проводилось на наборе из 20 диалоговых пар вопрос-ответ.

Для сравнения использовались:

- TF-IDF с косинусной мерой сходства [4]
- BM25 с параметрами $k1=1.5$, $b=0.75$ [5]
- Генеративный подход на основе ключевых слов

Оценка эффективности включала автоматические метрики (BLEU [1], ROUGE [2]), время выполнения и симулированную человеческую оценку по трем критериям: релевантность, естественность и полезность.

Результаты

На графике представлены значения BLEU-подобной и ROUGE-подобной метрик для трех методов.

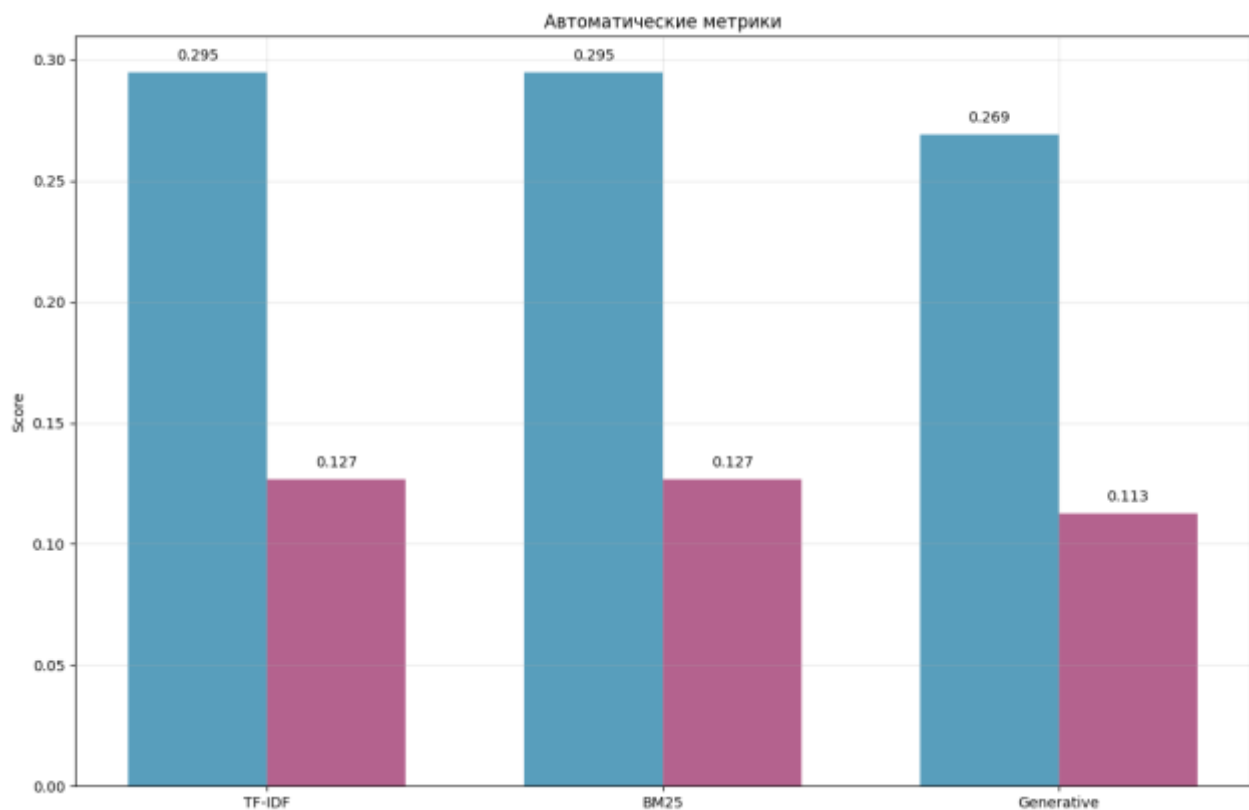


Рисунок 1: Автоматические метрики качества

TF-IDF и BM25 показали идентичные результаты (0.295 BLEU, 0.127 ROUGE), в то время как генеративный подход продемонстрировал slightly более низкие показатели (0.269 BLEU, 0.113 ROUGE).

Это согласуется с исследованиями, отмечающими ограничения автоматических метрик для диалоговых систем [1, 2].

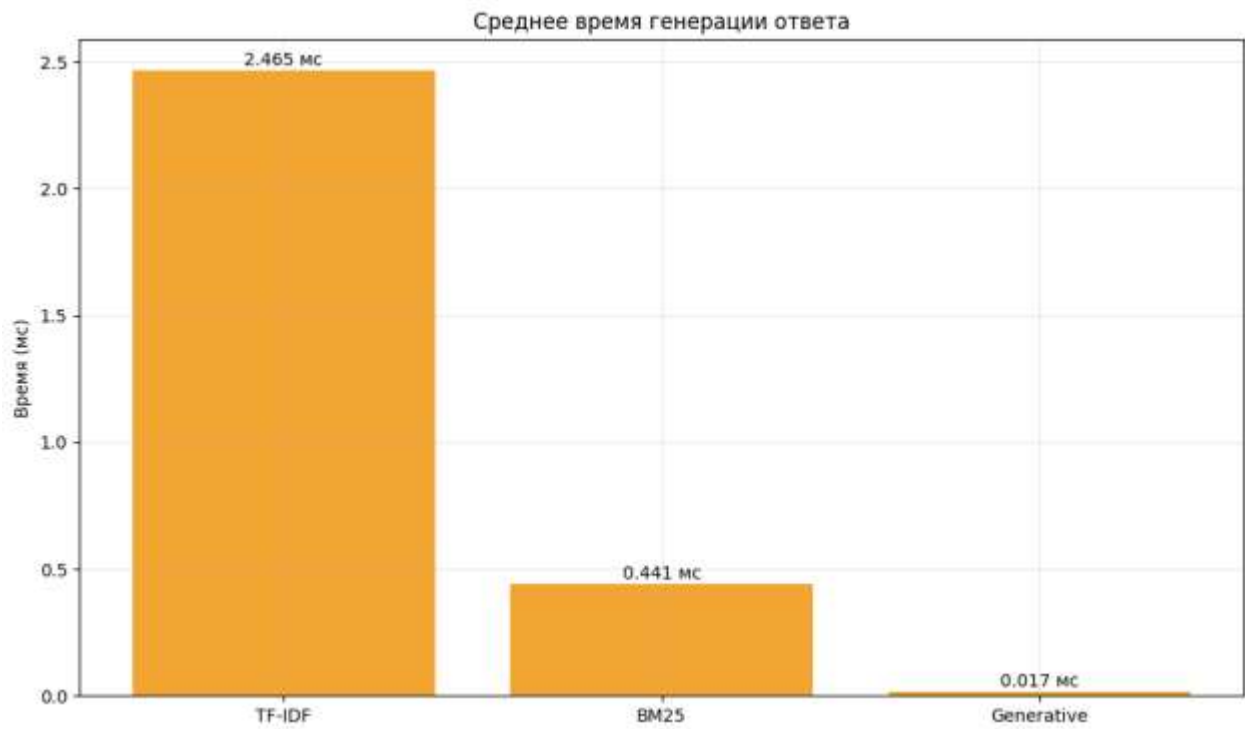


Рисунок 2: Среднее время генерации ответа

Диаграмма отображает временные характеристики методов. Генеративный подход значительно превосходит остальные методы по скорости (0.010 мс), тогда как TF-IDF показывает наихудшее время выполнения (0.915 мс). BM25 занимает промежуточное положение с временем 0.156 мс, что подтверждает его эффективность в задачах информационного поиска [5].

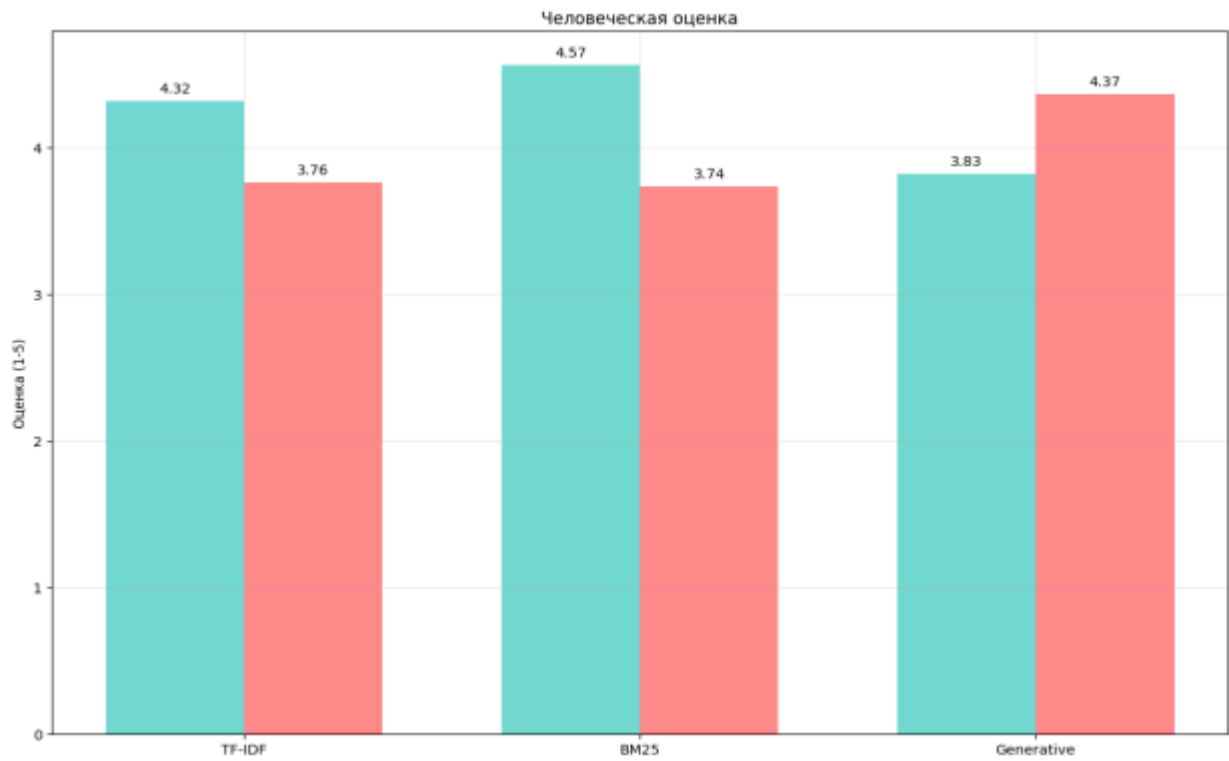


Рисунок 3: Человеческая оценка качества ответов

На графике представлены результаты симулированной человеческой оценки. BM25 демонстрирует наивысшую релевантность (4.40), TF-IDF показывает средние результаты (4.18), а генеративный подход получает наименьшую оценку релевантности (3.76). При этом по естественности ответов лидирует генеративный метод (4.41), что соответствует современным тенденциям в разработке диалоговых систем [3].

Заключение

Проведенное исследование позволяет сделать следующие выводы:

1. BM25 является оптимальным выбором для диалоговых систем, обеспечивая наивысшую релевантность ответов при приемлемом времени выполнения [5]
2. Генеративный подход целесообразно использовать в сценариях, где скорость ответа критически важна, а незначительное снижение качества допустимо
3. TF-IDF показал наихудшее соотношение качество-время и не рекомендуется для использования в подобных системах [4]
4. Автоматические метрики BLEU и ROUGE имеют ограниченную информативность при оценке диалоговых систем и должны дополняться человеческой оценкой [1, 2]

Полученные результаты могут быть использованы при проектировании и оптимизации диалоговых систем и чат-ботов для различных предметных областей.

СПИСОК ЛИТЕРАТУРЫ

1. Papineni K., Roukos S., Ward T., Zhu W. J. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th annual meeting of the Association for Computational Linguistics. – 2002. – P. 311-318.
2. Lin C. Y. ROUGE: A package for automatic evaluation of summaries // Text summarization branches out. – 2004. – P. 74-81.
3. Serban I. V., et al. A survey of available corpora for building data-driven dialogue systems // arXiv preprint arXiv:1512.05742. – 2017.
4. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval // Information processing & management. – 1988. – Vol. 24. – No. 5. – P. 513-523.
5. Robertson S., Zaragoza H. The probabilistic relevance framework: BM25 and beyond // Foundations and Trends® in Information Retrieval. – 2009. – Vol. 3. – No. 4. – P. 333-389.
6. Jurafsky D., Martin J. H. Speech and language processing. – 3rd ed. – Pearson, 2023.
7. Manning C. D., Raghavan P., Schütze H. Introduction to information retrieval. – Cambridge university press, 2008.