УДК 004.93'1:004.85

Быстрицкий Никита Александрович,
магистр, Уральский федеральный университет имени первого Президента России Б. Н. Ельцина, г. Екатеринбург

СРАВНИТЕЛЬНЫЙ АНАЛИЗ OPEN-SOURCE РЕШЕНИЙ ДЛЯ РАСПОЗНАВАНИЯ МАТЕМАТИЧЕСКИХ ФОРМУЛ В ФОРМАТЕ LATEX

Аннотация: *В статье представлен сравнительный анализ семи актуальных open-source инструментов оптического распознавания символов (OCR) применительно к задаче распознавания математических формул с преобразованием в формат LaTeX. Рассматриваются системы Pix2Text, PaddleOCR (PP-FormulaNet), Tesseract OCR, EasyOCR, UniMERNet, Mathpix Snip и Nougat. Для оценки качества распознавания применяются три метрики: Character Error Rate (CER), Exact Match и BLEU. Тестирование проводится на трёх независимых наборах данных: Im2LaTeX-100K (печатные формулы), HME100K (рукописные формулы) и синтетическом датасете. Установлено, что все рассмотренные решения демонстрируют значительное снижение качества на синтетически сгенерированных изображениях формул. На основании результатов тестирования обоснован выбор Pix2Text и PP-FormulaNet в качестве базовых кандидатов для дообучения.*

Ключевые слова: *оптическое распознавание символов, математические формулы, LaTeX, машинное обучение, синтетические данные, сравнительный анализ, трансформерные модели.*

Bystritsky Nikita Aleksandrovich,
Master's student, Ural Federal University Yekaterinburg

COMPARATIVE ANALYSIS OF OPEN-SOURCE SOLUTIONS FOR MATHEMATICAL FORMULA RECOGNITION IN LATEX FORMAT

Abstract: *The article presents a comparative analysis of seven current open-source optical character recognition (OCR) tools for the task of recognizing mathematical formulas and converting them to LaTeX format. The study covers Pix2Text, PaddleOCR (PP-FormulaNet), Tesseract OCR, EasyOCR, UniMERNet, Mathpix Snip, and Nougat. Three metrics are used for quality evaluation: Character Error Rate (CER), Exact Match, and BLEU. Testing is conducted on three independent datasets: Im2LaTeX-100K (printed formulas), HME100K (handwritten formulas), and a synthetic dataset. It is established that all considered solutions demonstrate a significant quality degradation on synthetically generated formula images. Based on the test results, Pix2Text and PP-FormulaNet are justified as baseline candidates for fine-tuning.*

Keywords: *optical character recognition, mathematical formulas, LaTeX, machine learning, synthetic data, comparative analysis, transformer models.*

Введение

Оптическое распознавание символов (OCR) является одной из ключевых технологий автоматизации обработки документов. Распознавание математических формул представляет особую сложность ввиду двумерной структуры нотации: показатели степени, дроби, интегралы и матрицы образуют иерархические конструкции, недоступные для стандартных линейных OCR-методов [1].

Стандартом для представления математических формул в научном сообществе является язык разметки LaTeX [2], что делает задачу автоматического преобразования изображений формул в LaTeX-код особенно востребованной. На сегодняшний день существует ряд open-source решений, претендующих на решение данной задачи, однако систематический

сравнительный анализ их эффективности на единой методологической основе в литературе представлен недостаточно.

Цель настоящей работы – провести сравнительный анализ семи актуальных open-source OCR-инструментов, ориентированных на распознавание математических формул, с применением унифицированного набора метрик на трёх независимых тестовых наборах данных.

Результаты

Обзор рассматриваемых инструментов

Для анализа были отобраны семь open-source решений. Tesseract OCR [3] – одна из наиболее зрелых библиотек распознавания текста, поддерживающая свыше 100 языков; однако её архитектура ориентирована на линейные текстовые строки, что принципиально ограничивает применимость к формульному контенту. PaddleOCR [4] включает специализированную модель PP-FormulaNet, предназначенную для распознавания математических формул; доступна в конфигурациях plus-S и plus-M, различающихся размером и вычислительными требованиями. EasyOCR оптимизирована для текста, не формул; коммерческая лицензия обязательна при годовом обороте свыше двух миллионов долларов США. UniMERNet [5] – специализированная модель для распознавания математических выражений в реальных сценариях, ориентированная преимущественно на латинскую нотацию. Mathpix Snip – высокоточный коммерческий инструмент с закрытым исходным кодом, исключающий дообучение. Nougat [6] – модель для извлечения структурированного текста из PDF-файлов научных статей, не адаптированная для обработки отдельных изображений формул. Pix2Text [7] – open-source альтернатива Mathpix, поддерживающая свыше 80 языков.

Методология тестирования

Для оценки качества распознавания применялись три метрики [8]. Character Error Rate (CER) – нормированное расстояние Левенштейна на уровне символов; значение 0 соответствует идеальному распознаванию,

меньше – лучше. Exact Match (EM) – доля изображений, для которых предсказание совпадает с эталоном посимвольно. BLEU [9] – взвешенное геометрическое среднее точностей совпадений n-граммов; значение выше – лучше.

Тестирование проводилось на трёх наборах данных. Im2LaTeX-100K [10] содержит свыше 100 000 пар «изображение – LaTeX-разметка» из реальных научных публикаций репозитория arXiv. HME100K включает 100 000 изображений рукописных математических выражений, написанных различными авторами. Синтетический датасет сформирован с применением генератора изображений с нестандартными шрифтами, фонами и геометрическими искажениями.

Результаты тестирования

Результаты сравнительного тестирования по двум метрикам – CER и BLEU – на трёх наборах данных представлены в таблице 1.

Таблица 1.

Результаты тестирования моделей (CER и BLEU)

Модель	**CER Im2LaTeX**	**BLEU Im2LaTeX**	**CER HME100K**	**BLEU HME100K**	**BLEU Синтетика**	**EM Синтетика**
Pix2Text	0,15	0,85	0,10	0,93	0,29	0,02
PP-FormulaNet-S	0,20	0,79	0,13	0,85	0,11	0,004
PP-FormulaNet-M	0,19	0,81	0,11	0,90	0,13	0,007
Tesseract OCR	–	–	–	–	–	–
EasyOCR	–	–	–	–	–	–

Tesseract OCR и EasyOCR не тестировались ввиду принципиальной неприспособленности к формульному контенту.

Анализ результатов

На датасете Im2LaTeX-100K наилучшие результаты показала модель Pix2Text: CER = 0,15, BLEU = 0,85. Модели PP-FormulaNet-plus-S и PP-FormulaNet-plus-M продемонстрировали сопоставимые результаты (BLEU 0,79 и 0,81 соответственно), незначительно уступая Pix2Text. На рукописных данных HME100K Pix2Text также сохраняет лидерство: BLEU = 0,93, что свидетельствует о высокой обобщающей способности модели за пределами типичного обучающего домена.

Принципиально иная картина наблюдается на синтетическом датасете. Все рассмотренные модели демонстрируют значительное падение качества: Pix2Text – BLEU = 0,29, PP-FormulaNet-plus-S – BLEU = 0,11, PP-FormulaNet-plus-M – BLEU = 0,13. Данный эффект объясняется доменным смещением (domain shift): модели, обученные на изображениях из реальных научных публикаций, не способны адекватно обрабатывать синтетически сгенерированные изображения с нестандартными шрифтами, фонами и геометрическими искажениями.

Tesseract OCR и EasyOCR были исключены из численного тестирования ввиду их архитектурной ориентированности на линейный текст. Mathpix Snip и Nougat исключены по причине закрытости исходного кода (Mathpix) и ориентированности на полные документы, а не отдельные изображения формул (Nougat).

UniMERNet при очень высокой точности имеет ограниченную применимость в русскоязычном контексте ввиду акцента на латинскую нотацию при обучении.

Заключение

Проведённый сравнительный анализ позволил сформулировать следующие выводы. Среди рассмотренных open-source решений наилучшее качество распознавания печатных и рукописных формул демонстрирует Pix2Text. Модель PP-FormulaNet в конфигурации plus-M является ближайшим конкурентом и превосходит plus-S по ряду метрик. Все рассмотренные

решения существенно теряют качество на синтетических данных, что указывает на актуальность разработки специализированных методов обучения, устойчивых к доменному смещению. Результаты тестирования подтверждают необходимость применения специализированных метрик (CER, BLEU) вместо общих метрик точности для объективной оценки качества распознавания LaTeX-формул. Полученные результаты могут служить основой для выбора базовой модели при разработке специализированных OCR-модулей для распознавания математического контента.

Использованные источники:

1. Zanibbi R., Blostein D. Recognition and retrieval of mathematical expressions // International Journal of Document Analysis and Recognition. 2012. Vol. 15, № 4. P. 331–357.

2. Документация системы LaTeX2e [Электронный ресурс] // The LaTeX Project. URL: https://www.latex-project.org/help/documentation/ (дата обращения: 05.05.2026).

3. Smith R. An overview of the Tesseract OCR engine // Proceedings of the 9th International Conference on Document Analysis and Recognition. IEEE, 2007. Vol. 2. P. 629–633.

4. Du Y. et al. PP-OCR: a practical ultra lightweight OCR system // arXiv preprint arXiv:2009.09941. 2020. URL: https://arxiv.org/abs/2009.09941 (дата обращения: 06.05.2026).

5. Wang B. et al. UniMERNet: a universal network for real-world mathematical expression recognition // arXiv preprint arXiv:2404.15254. 2024. URL: https://arxiv.org/abs/2404.15254 (дата обращения: 06.05.2026).

6. Blecher L. et al. Nougat: neural optical understanding for academic documents // arXiv preprint arXiv:2308.13418. 2023. URL: https://arxiv.org/abs/2308.13418 (дата обращения: 07.05.2026).

7. Pix2Text: открытая альтернатива Mathpix [Электронный ресурс] / Breezedeus. URL: https://github.com/breezedeus/Pix2Text (дата обращения: 07.05.2026).

8. Бобрышов А. П. Оценка качества работы библиотек оптического распознавания символов // Компетентность / Competency (Russia). 2025. № 7. URL: https://cyberleninka.ru/article/n/otsenka-kachestva-raboty-bibliotek-opticheskogo-raspoznavaniya-simvolov (дата обращения: 08.05.2026).

9. Papineni K. et al. BLEU: a method for automatic evaluation of machine translation // Proceedings of the 40th Annual Meeting of the ACL. 2002. P. 311–318. URL: https://aclanthology.org/P02-1040/ (дата обращения: 08.05.2026).

10. Deng Y. et al. What you get is what you see: a visual markup decompiler // arXiv preprint arXiv:1609.04938. 2016. URL: https://arxiv.org/abs/1609.04938 (дата обращения: 10.05.2026).