ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2023, том 514, № 2, с. 297–307

УДК 004.(89+93)

ПОИСК ТЕКСТОВЫХ ЗАИМСТВОВАНИЙ В РУКОПИСНЫХ ТЕКСТАХ

© 2023 г. А. В. Грабовой[1,2,*], М. С. Каприелова[1,2,3,], А. С. Кильдяков[1,***], И. О. Потяшин[1,****], Т. Б. Сейил[1,*****], Е. Л. Финогеев[1,******], Ю. В. Чехович[1,3,*******]**

Представлено академиком РАН А.Л. Семеновым
Поступило 02.09.2023 г.
После доработки 15.09.2023 г.
Принято к публикации 18.10.2023 г.

Поиск заимствований в учебных работах становится в последнее время все более актуальной задачей. Повышение популярности онлайн-образования, активная экспансия онлайн-платформ, ориентированных на среднее образование, формируют потребность в инструменте, способном проверять на заимствования рукописные работы школьников. Существующие подходы к поиску рукописных заимствований не подходят для быстрой проверки значительного количества работ по большими коллекциям потенциальных источников. Это существенно ограничивает их применимость. Кроме того, на практике требуется обрабатывать изображения текстовых страниц посредственного качества, выполненные, как правило, с помощью фотокамер мобильных телефонов. Мы предлагаем метод, который позволяет выявлять заимствованные фрагменты текста в документах, представленных в виде изображений (фотографий) рукописных текстов, при сопоставлении с большими коллекциями источников. Метод включает в себя три этапа: распознавание символов рукописного текста, поиск кандидатов и последующий точный поиск источника заимствований. В работе приведены результаты экспериментов по оценке качества и производительности разработанной системы. Полнота поиска заимствований в рукописных документах достигает 83.3% при обработке изображений высокого качества и 77.4% при обработке изображений худшего качества. Время выполнения поиска для одного документа по коллекции источников из 100000 документов составляет в среднем 3.2 с при использовании CPU. Результаты показали, что созданная нами система может быть масштабирована и использована для промышленных задач, требующих быстрой проверки сотен тысяч школьных сочинений по большому количеству потенциальных источников заимствований. Все эксперименты проводились на открытом наборе данных HWR200.

Ключевые слова: оптическое распознавание символов, рукописный текст, поиск текстовых заимствований, компьютерное зрение, распознавание рукописного текста

DOI: 10.31857/S2686954323601720, **EDN:** XXIYGC

1. ВВЕДЕНИЕ

Задача выявления плагиата в учебных работах в течение последних десятилетий приобрела высокую актуальность. Современные информационные технологии существенным образом упростили поиск текста и его копирование. Значительный рост количества случаев плагиата в учебных работах оказался одним из негативных результатов технологического развития [1, 2]. В настоящее время значительная часть вузов и научных организаций использует системы обнаружения заимствований. При этом исследования показывают, что привычку копировать чужие тексты без ссылки на источник или копировать готовые работы из интернета обучающиеся приобретают еще в школе и “приносят с собой” в вузы, а затем и в свою профессиональную деятельность [3]. Существенным отличием среднего образования от высшего является использование массовое рукописных форм учебных работ. Кроме того, в последнее время активное развитие онлайн-образования приводит к тому, что зачастую школьникам предлагается отправлять преподавателю отсканированные или сфотографированные рукописные работы. Очевидно, что в таких условиях невозможно обеспечивать прежний уровень

[1]*Компания Антиплагиат, Москва, Россия*
[2]*Московский физико-технический институт, Москва, Россия*
[3]*Федеральный исследовательский центр “Информатика и управление” Российской академии наук, Москва, Россия*
**E-mail: grabovoy@ap-team.ru*
***E-mail: kaprielova@ap-team.ru*
****E-mail: kildyakov@ap-team.ru*
*****E-mail: potyashin@ap-team.ru*
******E-mail: seilov@ap-team.ru*
*******E-mail: finogeev@ap-team.ru*
********E-mail: chehovich@ap-team.ru*

контроля преподавателей за процессом подготовки работ. В результате некоторые ученики могут попросту переписывать (возможно, частично) работы, выполненные другим человеком. Источниками для списывания могут также служить материалы других школьников, которые выкладывают готовые работы в общий доступ в интернете. У проверяющих преподавателей попросту нет возможности проверять каждую работу школьника на оригинальность, сравнивая ее на наличие заимствований со всеми возможными источниками. При этом проверка результатов государственных экзаменов в РФ в соответствии с ФГОС требует обязательной проверки работ на заимствования. В масштабах страны возникает необходимость обработать сотни тысяч работ за ограниченное время. Существование автоматической проверки на наличие заимствований значительно облегчит работу проверяющих. В настоящей работе мы предлагаем систему поиска заимствований в рукописных текстах. Система решает проблему поиска заимствований в рукописных работах в три этапа. Первый — преобразование изображения в последовательность символов с помощью оптического распознавания. Второй этап — поиск заимствований — включает в себя поиск кандидатов, который сужает число возможных источников для заимствований. Третий этап — точное сравнение, оставляющее из небольшого количества кандидатов наиболее вероятные. Такой подход позволяет не только работать с языками, качество распознавания рукописного текста для которых недостаточно высокое, но и масштабировать задачу поиска заимствований на большие текстовые коллекции.

2. ОБЗОР ЛИТЕРАТУРЫ

В литературе описывается несколько подходов к решению задачи поиска заимствований в текстах рукописных работ. Их можно разделить на две группы: использующие методы распознавания рукописного текста и основывающиеся на других принципах. Примером второго подхода является [4], где авторы предлагают сравнивать векторные представления отдельных слов, полученные с помощью сверточной нейросети. Недостаток этого метода заключается в том, что для обучения модели требуют большое количество размеченных данных. Авторы [5] предложили ставить в соответствие тексту нормированную последовательность длин слов. Длины слов оцениваются исходя из размеров прямоугольников, ограничивающих рукописные слова в изображении. Этот метод не требует применения оптического распознавания символов, но имеет целый ряд недостатков, среди которых гипотеза о небольшом (от одного до трех) максимальном количестве источников заимствования, нестойкость к переносам слов, линейная от размера коллекции источников зависимость вычислительных затрат на обработку каждого документа.

Отдельно следует выделить подходы, в которых система оптического распознавания рукописного текста совмещена с системой поиска заимствований [6]. Их популярность связана с большим количеством практических исследований в области распознавания рукописного текста [7–9]. Однако решения этого типа имеют ряд недостатков. Во-первых, они также требуют разметки на этапе обучения. Во-вторых, для достижения хорошего качества распознавания используются “тяжелые” модели, не слишком эффективные при работе с большими коллекциями данных. Кроме того, задача оптического распознавания рукописного текста чувствительна к языку распознаваемого текста. По этой причине методы решения задачи поиска заимствований в рукописных текстах, основанные только на распознавании рукописного текста, не очень эффективны для языков с недостаточным объемом размеченных текстов, находящихся в открытом доступе. Примером такого языка является русский язык. Так, одна из лучших открытых моделей распознавания рукописного текста на русском языке StackMix-OCR [10], обученная на датасете HKR [11], хоть и дает на нем хорошее качество, но при применении на обычных текстах вне датасета не может дать такого же результата, что можно видеть на рис. 1.

3. ПОСТАНОВКА ЗАДАЧИ

В общем случае задачу поиска заимствований в рукописных текстах можно сформулировать следующим образом.

Существует набор документов-оригиналов $D_{orig} = \{d_1^o, \ldots, d_n^o\}$ и множество документов-запросов $D_q = \{d_1^q, \ldots, d_m^q\}$, наличие заимствований в которых требуется определить.

Документы-оригиналы и документы-запросы представимы в виде конкатенации фрагментов

$$d_i^o = l_{i_1}^o \sqcup \ldots \sqcup l_{i_k}^o,$$

$$d_j^q = l_{j_1}^q \sqcup \ldots \sqcup l_{j_p}^q.$$

Пусть задана выборка

$$D = \{(d_x^o, d_x^q), RW_x\}_{x=1}^X,$$

где каждой паре документов (d_x^o, d_x^q) сопоставлен список пар фрагментов

$$RW_x = \{(l_{x_1}^o, l_{x_1}^q), \ldots, (l_{x_w}^o, l_{x_w}^q)\}.$$

Для каждой пары фрагметнов известно, что фрагмент $l_{x_a}^q$ является заимствованием фрагмента

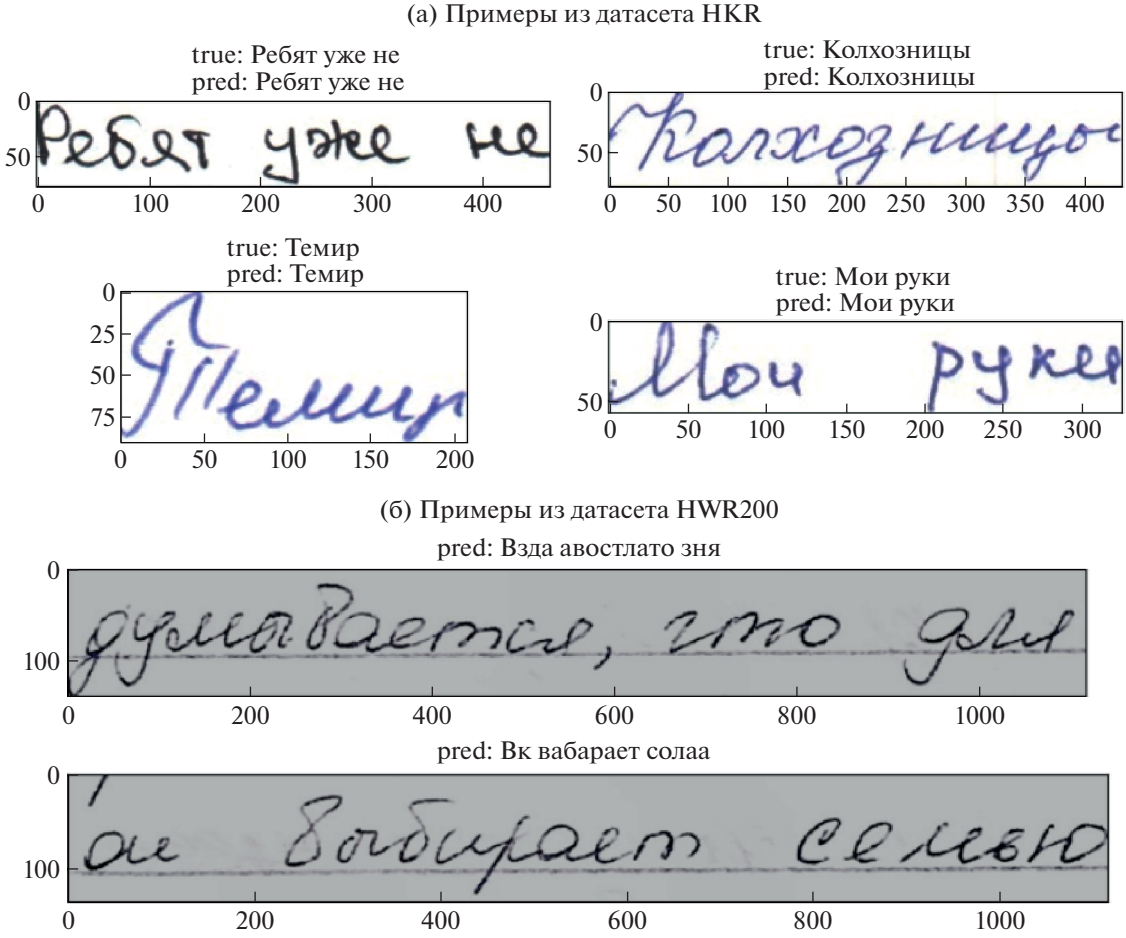


Рис. 1. Пример ипользования модели StackMix-OCR, обученной на датасете HKR.

$l^o_{x_a}$. Необходимо найти функцию $F = f \circ g \circ h$, которая является суперпозицией функции распознавания рукописного теста

$$h = D \to D^c,$$

где D^c — множество документов, содержащих распознанный текст, функции поиска кандидатов

$$g = (d_s^{qc}, D^c_{orig})_{d_s^{qc} \in D_q^c} \to D_{orig}^{retrieved_s} \subset D_{orig}{}^c,$$

где $D_{orig}^{retrieved_s}$ — множество документов-кандидатов, и функции точного сравнения

$$f = (d_s^{qc}, D_{orig}^{retrieved_s})_{d_s^{qc}} \in D_q^c \to RW_s.$$

Качество модели оценивается функцией

$$Recall@1 = \frac{|(\cup_{x=1}^{X} RW_x) \cap (\cup_{j=1}^{J} \widehat{RW_j})|}{|\cup_{x=1}^{X} RW_x|}.$$

Искомая функция $\hat{F}$ должна максимизировать метрику полноты

$$\hat{F} = \arg\max_{F \in \mathcal{F}}(Recall@1(F, D)),$$

где $\mathcal{F}$ — множество рассматриваемых функций.

4. МЕТОД

В разработанной системе поиска заимствований в рукописных текстах каждый рукописный документ проходит через несколько стадий обработки.

Сначала производится определение ориентации документа и разбиение документа на строки. Эта операция необходима, так как алгоритм распознавания рукописного текста работает с каждой строкой отдельно. Стадия предобработки реализована следующим образом: сначала производится сглаживание, перевод в черно-белый формат, затем производится увеличение контрастности изображения с помощью применения адаптивного порога. Таким образом, может быть получена маска, выделяющая текст. Далее матрица изображения суммируется по строкам и получается сигнал. Исходя из формы полученной последова-

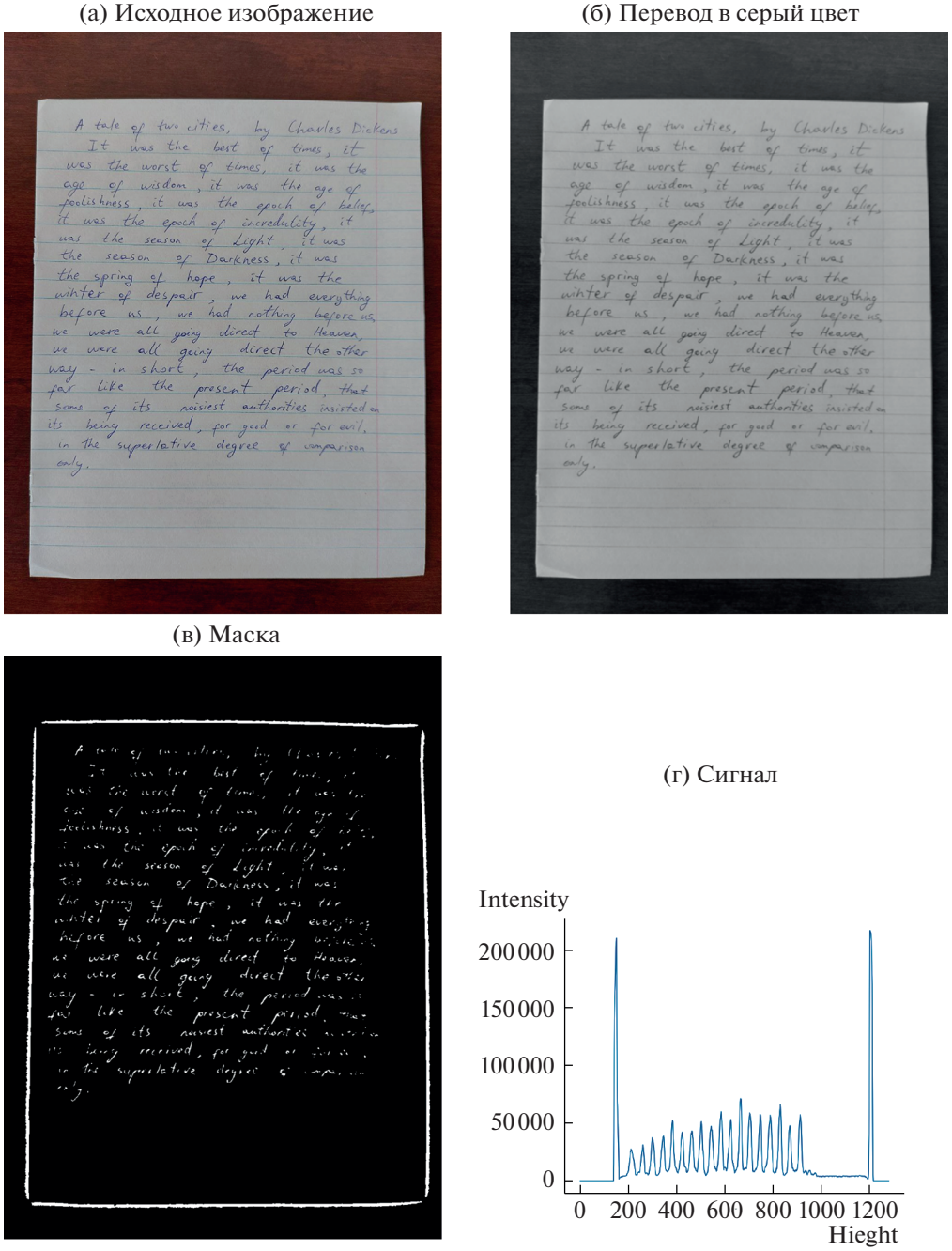


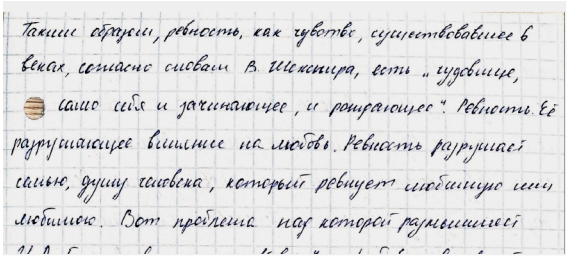
Рис. 2. Получение сигнала из изображения.

тельности, определяется ориентация страницы. Для корректно ориентированного изображения характерен профиль с несколькими ярко выделенными пиками, соответствующими координатам строк. Пример получения сигнала представлен на рис. 2.

Для разбиения на строки из сигнала извлекаются максимумы и минимумы, которые позволяют локализовать строки текста на странице. Максимумы сигнала соответствуют сегментам изображения, где расположен текст, а минимумы позволяют определить участки, где текста нет. Таким образом, могут быть обнаружены строки текста на странице и изображение делится по минимумам сигнала.

Следующая стадия — распознавание рукописного текста. Для перевода рукописного текста в машиночитаемый был использован метод глубокого обучения в области оптического распознавания рукописного текста. Модель типа кодиров-

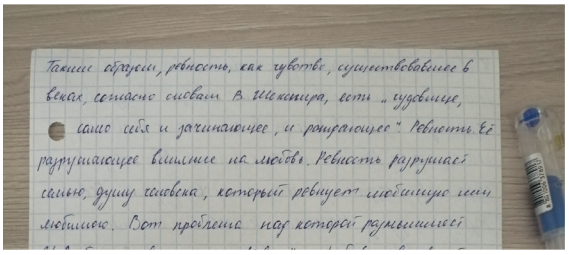
(а) Скан



Результат работы модели

```
Пакиии образем, ребность, как чувотва,
сущетвовавшее в венах, сожасно словам В.
Иисненира, есть, гудовице, сомо себя и
зачинающее, и ротрающес" Ревность. Сё
разрусшакщее влилние на любовь. Невность
разрушаст салью, душу человека, который
ревнует любимую мии любимого. Вот
проблета над которой разньишичест
```

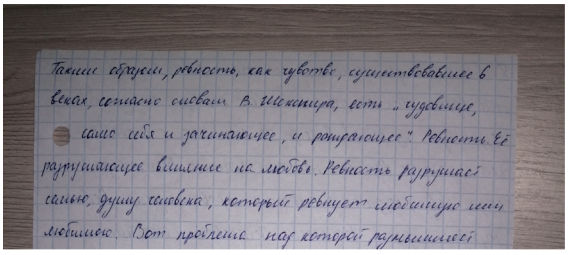
(б) Светлое фото



Результат работы модели

```
а ори, ват и узвов, дения венах, сотасно
словам В. Иисненира, есть гудовице, сомо
себя и зачинающее, и ротдающес". Ревность.
С разтушакце влияние на любовь. Невность
разрушаст семью, думу человека, который
ровнует любимую нии любимого.
Вот проблета над которой разньшичест
```

(в) Темное фото



Результат работы модели

```
Дли срори ват увов, учанел венах, сотасно
словам В. чискнира, есть, гудовице, сомо
себя и зачинающее, и ротрающес" Ревность.
С разтушакце влилние на любовь. Невность
разрушаст самью, душу человска, который
ревнует любимую иии любимого. Вот проблета
над которой размьшшичест
```

Рис. 3. Пример применения используемой OCR-модели для изображений из HWR200.

щик-декодировщик была разработана по мотивам работы [7]. Стоит отметить, что для задачи распознавания рукописного текста на русском языке существует не так много датасетов для обучения. Также важно указать, что для применения на большой коллекции документов модель должна отрабатывать достаточно быстро, поэтому не должна быть тяжелой. Эти два обстоятельства не позволяют получить такое качество распознавания рукописных текстов, чтобы искать заимствования исключительно на основании распознанного текста. Пример использования OCR-модели на датасете HWR200 [12], изображение в котором можно считать максимально приближенным к реальным, можно увидеть на рис. 3. В связи с этим были внедрены еще несколько этапов обработки полученного текста: составление символьных n-грамм, поиск по индексу n-грамм, сегментация текста, векторизация и сравнение векторных представлений. Отдельно подчеркнем, что решение задачи для распознавания рукописного теста, обученное на другом языке, не сможет эффективно работать на русскоязычных данных.

Затем производится поиск источников заимствований. Эту часть системы можно условно разделить на два этапа: поиск кандидатов и точное сопоставление кандидатов и документа-запроса. Для эффективного поиска текстов-кандидатов был разработан подход, основанный на использовании индекса шинглов, предложенного в [13] и [14]. Каждому документу D в соответствие ставится $S(D,w)$ — набор шинглов размера w. Тогда при фиксированном размере шингла мера совпадения документов может быть записана следующим образом:

$$R(D_{orig}, D_q) = \frac{|S(D_{orig}) \cap S(D_q)|}{|S(D_{orig}) \cup S(D_q)|}.$$

Для применеия этого подхода производится предобработка текста. Сначала слова делятся на n-граммы. По результатам экспериментов оптимальным количеством символов было признано 2 (два). N-граммы такого вида называются биграммами. На следующем этапе последовательность биграмм представляется в виде шинглов — перекрывающихся подпоследовательностей биграмм некоторой длины k. Например, первый шингл составляют биграммы с первого по k-е, второй шингл — со второго по $(k+1)$-е и так далее до конца текста. Значение k, как правило, выбирает-

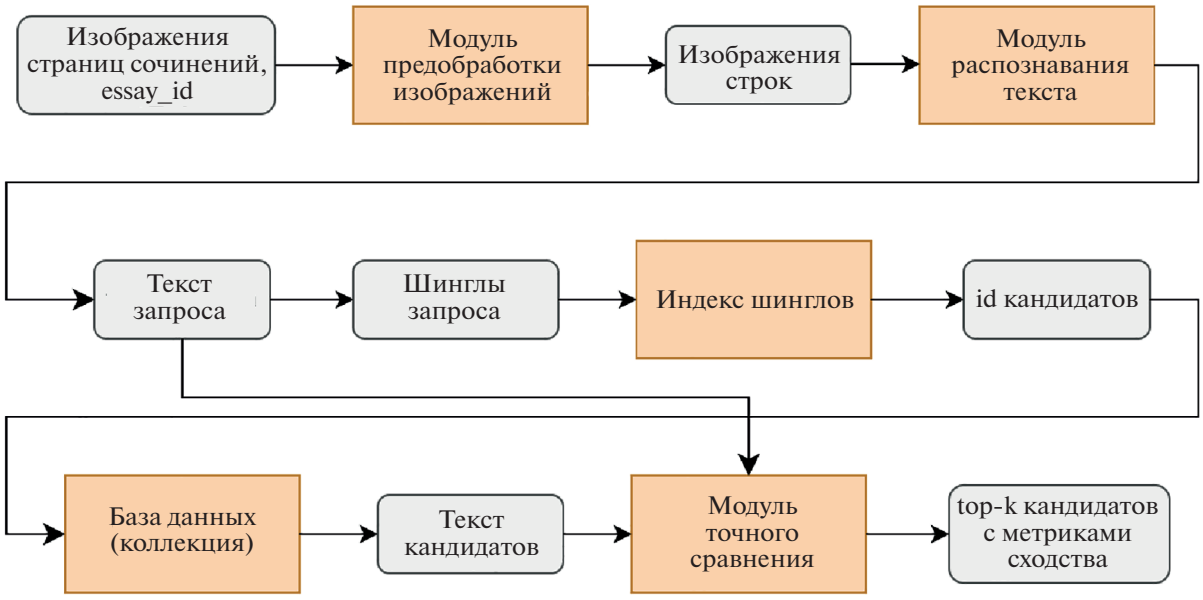


Рис. 4. Схема алгоритма поиска рукописных текстов в коллекции.

ся небольшим. Эксперименты показали, что оптимальным значением k является 2 (два). Далее каждый шингл хэшируется. Полученные значения хэшей шинглов вместе с идентификатором текста и позицией этого шингла в тексте записываются в специальный файл, который называется индексом. Составление такого файла позволяет осуществлять поиск кандидатов с логарифмической сложностью.

Для выполнения поиска проверяемый текст проходит аналогичные процедуры: предобработку и шинглирование. Затем производится бинарный поиск в индексном файле полученных значений. Если находятся совпадения, то найденные идентификаторы текстов попадают в список текстов-кандидатов. Для каждого текста-кандидата производится подсчет количества совпадений значений хэшей со значениями хэшей шинглов проверяемого документа.

Далее определяются точные границы совпавшего текста по отношению к строкам, а не к их шинглам. Для достижения этой цели производится сопоставление каждой строки-кандидата с проверяемой строкой с помощью суффиксного массива [15]. Мерой совпадения p проверяемой строки со строкой-кандидатом считается суммарная длина фрагментов, совпавших в обоих строках, отнесенная к длине проверяемого документа. Суммарная длина фрагментов достигает максимума, равного единице, если весь проверяемый текст строки содержится в тексте-кандидате.

Затем производится векторизация полученных фрагментов по строкам или областям – нескольким объединенным строкам. Для этого векторизуются строки текста-запроса и текста-кандидата. Перевод в векторное представление осуществляется с помощью библиотеки sklearn [16].

Финальная стадия – точное сопоставление кандидатов производится путем сравнения векторных представлений. Оно реализуется с помощью составления индекса по одному из изображений и поиска совпадающих строк из другого. Индекс строится с помощью библиотеки annoy, позволяющей производить поиск ближайших соседей. Затем на этом же этапе подсчитывается близость с помощью средств библиотеки для сравнения последовательностей difflib методом SequenceMatcher.

Общую схему алгоритма поиска можно увидеть на рис. 4. Желтые блоки на диаграмме соответствуют модулям, организованным в виде отдельных сервисов. Модуль точного сравнения включает в себя векторизацию кандидатов и сравнение векторных представлений запроса и кандидатов. Доступ к каждому сервису может быть параллельным, индекс шинглов и база данных с коллекцией позволяют одновременное считывание данных.

5. ОПИСАНИЕ НАБОРОВ ДАННЫХ

В открытом доступе существует относительно небольшое число наборов данных для обучения моделей распознавания рукописного текста.

Отличительными атрибутами датасетов для задачи можно назвать язык и природу текста, наличие разметки на уровне линий и слов, наличие полных страниц с текстом. Стандартными кол-

лекциями данных для проверки работы моделей распознавания рукописного текста считаются датасеты IAM [17], Bentham [18] и Read2016 [19].

Последняя версия датасета IAM [17] – это набор отсканированных англоязычных текстов, написанных 657 различными людьми в 1500 специальных формах для заполнения. Датасет содержит как полные страницы с текстом, так и тексты в виде изолированных предложений, строк и слов. Общее число слов в коллекции – 115320. Датасет Bentham [18] представляет собой отсканированные манускрипты английского философа Джереми Бентама. Коллекция состоит из более чем 6000 документов и 25000 страниц. Несмотря на больший объем данных по сравнению с предыдущим датасетом, Bentham [18] имеет существенный минус – наличие почерка лишь одного человека. Обученная на этом датасете модель может оказаться не приспособленной для распознавания почерка других людей. Датасет Read2016 [19] состоит из исторических документов, написанных на ранненововерхненемецком языке. Всего в наборе находится около 30000 страниц рукописей. Как и в других истрических датасетах, все изображения получены при хорошем освещении и в одинаковых условиях. Существует несколько причин того, что датасеты, состоящие из исторических документов, не очень подходят для решения задачи распознавания произвольного рукописного текста. Во-первых, язык текстов может содержать устаревшие слова, буквы или стили письма. Во-вторых, исторические документы в таких датасетах хорошо отсканированы или сфотографированы. В процессе работы модели на вход редко приходят документы схожего качества. В-третьих, в исторических документах значительная часть выборки может быть написана одним почерком. Это не позволяет их использовать для обучения модели, которая должна быть устойчива к изменению почерка и условиям оцифровки текста. Стоит отметить, что IAM [17], Bentham [18] и Read2016 [19] содержат тексты, написанные на латинице.

Существуют датасеты с текстами, написанными на кириллице, в частности, на русском языке. Среди них — Digital Petr [20], коллекция отсканированных школьных тетрадей school_notebooks [21], sber-idpforms [22], HKR [11], HWR200 [12]. Датасет Digital Petr [20], как и коллекция Bentham, является набором отсканированных исторических документов, написанных одним человеком. Коллекция также содержит сегментацию текста на уровне линий. Всего датасет состоит из почти 10000 изображений и 50000 слов. Коллекция school_notebooks [21] состоит из 1857 изображений школьных тетрадей с полигональной разметкой на уровне слов, а также набора пар “изображение слова – транскрипция” для каждого слова. При составлении открытого датасета sber-idp-forms [22] асессорам предлагалось на специальных бланках вручную написать заданные слова или фразы. Данные содержат 5203 изображения прямоугольников с написанным текстом. Аналогичная идея с заполнением форм использовалась при создании коллекции HKR [11]. В этом датасете текст написан на русском и казахском языках (примерно 95% фраз написаны на русском языке). Датасет KOHTD [23] включает в себя свыше 140000 отсканированных изображений экзаменационных студенческих работ. Аннотация коллекции содержит информацию для сегментации текста по строкам. Отличительной чертой этого датасета является то, что он почти полностью (на 99%) состоит из текста, написанного на казахском языке. Несмотря на то что работы были написаны на расширенной кириллице (в отличие от IAM [17], Bentham [18]), этот датасет не является лучшим вариантом для обучения моделей, которые предполагается использовать на русскоязычных данных. Это обусловлено тем, что функция потерь CTC loss, которая используется при обучении алгоритмов оптического распознавания рукописных текстов, помогает модели запоминать наиболее вероятные серии символов. Для русского и казахского языков вероятности последовательностей букв сильно различаются, что приводит к посредственным результатам при тестировании моделей на русскоязычных данных.

Также существуют наборы данных в виде набора рукописных текстов, собранные для более специфических задач. Примером такого датасета является HWR200 [12], специально разработанный для поиска заимствований в рукописных текстах. Датасет построен на 35 текстовых фрагментах, из которых составлялись тексты с заимствованием. Каждый текст с заимствованием может содержать от одного до двух заимствованных непрерывных фрагментов из оригиналов (исходных 35 фрагментов). Между заимствованными отрывками, а также в начале и в конце текста, могут быть вставлены предложения из третьих источников. Также в наборе выделена группа “чистые сочинения”. “Чистые сочинения” – тексты, не имеющие текстов-заимствований в коллекции. Тексты этой группы нужны для проверки системы поиска заимствований на ложноположительные срабатывания. Тексты написаны разными почерками и доступны в трех версиях: фото в ярком освещении, фото в тусклом освещении и сканированный документ. Общее количество изображений в датасете более 30000. К сожалению, в этом датасете нет построчной разметки текста. Однако его можно использовать на этапе тестирования качества распознавания текста и в экспериментах по поиску заимствованных фрагментов в рукописях.

Таблица 1. Характеристики датасетов [12]

	HWR200	Bentham, Digital Petr	**IAM**	School notebooks	Sber-idp-forms, HKR
Тексты или фразы	тексты	тексты	фразы	тексты	фразы
Разметка на уровне строк	–	+	+	+	+
Различные почерки	–	–	+	+	+
Различные условия	+	–	–	–	–

6. ЭКСПЕРИМЕНТ

Для вычисления полноты поиска кандидатов был проведен эксперимент на данных из датасета HWR200 [12].

Данные для индексации состоят из изображений сочинений, включающих в себя "оригинальные" тексты и части текстов с заимствованием (всего 1035 текстов или 3798 изображений). Остальные 1650 текстов с заимствованием — это запросы для поиска (5742 изображений). Здесь и далее нужно иметь в виду, что в датасете каждое изображение представлено в трех вариантах: сканы, изображения со светлым освещением и с темным освещением. В эксперименте наборы данных для индексации и поиска — это все возможные варианты из девяти комбинаций пар способов оцифровки данных. Например: коллекция состоит из сканированных изображений, а запросы — это изображения со светлым фоном.

Модуль поиска кандидатов возвращает до десяти возможных кандидатов для каждого запроса. Считается, что источник заимствования для запроса был найден корректно в следующих двух случаях: если кандидат Z — это оригинал, на основе которого был сконструирован запрос, либо кандидат и запрос имеют общие пересечения друг с другом, т.е. были составлены из одного и того же первоначального документа.

Определим η как отображение, значением которого на запросе d_s^{qc} является множество $D_{orig}^{source_s}$ истинных источников запроса d_s^q

$$\eta = (d_s^{qc}, D_{orig}^c)_{d_s^{qc} \in D_q^c} \to D_{orig}^{source_s} \subset D_{orig}^c.$$

Метрикой для измерения качества модуля выбора кандидатов является *RecallCand*@10 — доля запросов, для которых среди кандидатов были найдены истинные источники.

$$RecallCand@10 = \frac{\sum_{i=1}^{|Q|} \mathbf{1}[|\eta(d_i^{qc}, D_{orig}^c) \cap g^{(10)}(d_i^{qc}, D_{orig}^c)| > 0]}{|Q|},$$

где Q — множество документов-запросов и $|g^{(k)}(d_i^{qc}, D_{orig}^c)| = k \; \forall i$.

После получения для каждого запроса небольшого списка кандидатов алгоритм вычисляет метрику сходства в каждой паре запрос-кандидат в модуле точного сравнения. Кандидаты со значением метрики, превышающим некоторое пороговое значение, считаются источниками, из которых были взяты заимствования для запроса. Выбор оптимального значения порога осуществляется следующим образом. Для каждого запроса из предыдущего этапа корректируется список кандидатов. В список кандидатов включаются все (до двух) оригинальные источники для текста-запроса, тексты с заимствованием, построенные на тех же оригиналах, что и запрос, а также тексты, не пересекающиеся с запросом. Всего 10 кандидатов для каждого запроса. Порог на сходство выбирается таким образом, чтобы полнота *Recall*@10 была максимальной при заданной верхней границе на FPR. Здесь FPR – это доля запросов, которые не имеют пересечений с текстами из коллекции, но для которых система нашла хотя бы один источник с заимствованиями.

$$FPR = \frac{\sum_{i=1}^{|Q_{FP}|} \mathbf{1}[|F(d_i^{qc}, D_{orig}^c)| > 0]}{|Q_{FP}|},$$

где Q_{FP} — множество документов, у которых нет источников в D_{orig}.

Таблица 2. Полнота модуля поиска кандидатов

Тип изображений коллекции	Тип изображений запросов	RecallCand@10
Светлые	Светлые	90.3%
Светлые	Темные	88.0%
Светлые	Сканы	89.2%
Темные	Светлые	88.3%
Темные	Темные	88.5%
Темные	Сканы	87.8%
Сканы	Светлые	87.2%
Сканы	Темные	87.1%
Сканы	Сканы	88.1%

Таблица 3. Полнота полного пайплайна поиска заимствований в рукописных текстах

Тип изображений коллекции	Тип изображений запросов	RecallFull@10	RecallFull@1	FPR
Светлые	Светлые	85.7%	83.3%	2.8%
Светлые	Темные	84.2%	80.1%	3.2%
Светлые	Сканы	84.6%	81.2%	3.7%
Темные	Светлые	84.3%	81.1%	3.5%
Темные	Темные	84.7%	81.5%	3.1%
Темные	Сканы	84.0%	80.7%	3.9%
Сканы	Светлые	81.2%	78.4%	3.9%
Сканы	Темные	80.3%	77.4%	4.1%
Сканы	Сканы	81.8%	78.8%	3.7%

Для тестирования качества работы полного пайплайна используются метрики *Recall@k* ($k = 1$, 10), а также FPR.

7. РЕЗУЛЬТАТЫ

Результаты модуля поиска кандидатов представлены в табл. 2. Интересно, что во всех случаях поиск имеет лучшие результаты, если запросы и база имеют изображения со схожим освещением. Одной из главных причин различия качества работы поиска при данных с разным освещением является разное качество работы модулей предобработки (разбиения на строки) и модуля распознавания текста на таких изображениях.

В табл. 3 приведены результаты поиска заимствований в рукописных текстах на полном пайплайне в зависимости от формата коллекции. Кандидаты проходят дополнительную фильтрацию в модуле точного сравнения. Различие в качестве поиска для различных способов оцифровки изображений коррелирует с результатами модуля поиска кандидатов.

Результаты получены при фиксированном пороге на схожесть, равном 0.2. Такой порог позволяет установить False Positive Rate на уровне 3–4%.

Для измерения скорости поиска была проиндексирована более крупная коллекция рукописных текстов (100000 одностраничных документов из закрытого набора данных).

Время поиска на CPU одного документа в такой коллекции в среднем составляет 3.2 с. Замеры времени производились на процессоре AMD Ryzen 9 3900XT 12-Core Processor с 12 ядрами и максимальной частотой 3800 МГц. Несмотря на то что в реальных условиях коллекции будут значительно больше (десятки, сотни миллионов документов), результаты измерений скорости поиска релевантны и для таких коллекций, т.к. время работы модулей предобработки, распознавания текстов и модуля точного сравнения не зависит от размера коллекции источников, а время поиска по индексу шинглов логарифмически зависит от числа проиндексированных документов. Кроме того, доступ к каждому сервису может быть параллельным, а индекс шинглов и база данных с коллекцией позволяют осуществлять одновременное считывание данных.

8. ЗАКЛЮЧЕНИЕ

В статье предложена система поиска заимствований в рукописных текстах. Разработанная система решает проблему в три этапа: оптическое распознавание рукописного текста, поиск кандидатов и точное сравнение кандидатов с документом. Представленный алгоритм позволяет находить источники заимствований с полнотой Recall@1=83.3% на изображениях с хорошим качеством и Recall@1=77.4% на изображениях с темным освещением и с небольшой долей ложноположительных срабатываний FPR на уровне 2–4%. Предлагаемая система эффективно работает с большими коллекциями, о чем свидетельствуют результаты замера скорости обработки документов. Время обработки одного документа составляет 3.2 с при поиске по коллекции, содержащей 100000 документов. Стоит отметить, что разработанная система может быть использована не только при поиске заимствований в рукописных работах по коллекции рукописных документов, но и по источникам в машинописном виде

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Работа выполнена при поддержке Фонда Содействия Инновациям (проект № 79068, заявка № ИИ-208298).

СПИСОК ЛИТЕРАТУРЫ

1. *Никитов А.В., Орчаков О.А., Чехович Ю.В.* Плагиат в работах студентов и аспирантов: проблема и методы противодействия.// Университетское управление: практика и анализ. 2012. № 5 (81). С. 61–68.
2. *Roig Miguel.* Avoiding plagiarism, self-plagiarism, and other questionable writing practices: A guide to ethical writing. 2011.
3. *Чехович Ю.В., Беленькая О.С.* Методика внедрения и использования электронных средств обнаружения заимствований в системе среднего образования // Информатика и образование. 2021. № 10 (329). С. 5–14. https://doi.org/10.32517/0234-0453-2021-36-10-5-14
4. *Praveen K., Jawahar C.V.* Matching handwritten document images. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14.
5. *Bakhteev O., Kuznetsova R., Khazov A., Ogaltsov A., Safin K., Gorlenko T., Suvorova M., Ivahnenko A., Botov P., Chekhovich Yu., Mottl V.* Near-duplicate handwritten document detection without text recognition. Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference "Dialogue". 2021. Issue 20. P. 47–57. https://doi.org/10.28995/2075-7182-2021-20-47-57
6. *Pandey Om, Gupta Ishan, Mishra Bhabani S.P.* A Robust Approach to Plagiarism Detection in Handwritten Documents. Advances in Visual Computing: 15th International Symposium, ISVC 2020, San Diego, CA, USA, October 5–7, 2020, Proceedings, Part II 15.
7. *Coquenet D., Chatelain C., Paquet Th.* End-to-end handwritten paragraph text recognition using a vertical attention network. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2022. V. 45 (1). P. 508–524.
8. *Rowtula V., Bhargavan V., Kumar M., Jawahar C.V.* Scaling handwritten student assessments with a document image workflow system. Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2018. P. 2307–2314.
9. *Voigtlaender P., Doetsch P., Ney H.* Handwriting recognition with large multidimensional long short-term memory recurrent neural networks, 15th international conference on frontiers in handwriting recognition (ICFHR). 2016. P. 228–233.
10. *Potanin M., Dimitrov D., Shonenkov A., Bataev V., Karachev D., Novopoltsev M., Chertok A.* Digital Peter: New Dataset, Competition and Handwriting Recognition Methods. The 6th International Workshop on Historical Document Imaging and Processing, Lausanne, Switzerland, 2021. P. 43–48.
11. *Nurseitov D., Bostanbekov K., Kurmankhojayev D., Alimova A., Abdallah A., Tolegenov R.* Handwritten Kazakh and Russian (HKR) database for text recognition. Multimedia Tools and Applications. 2021. V. 80. P. 33075–33097.
12. *Potyashin I., Kaprielova M., Chekhovich Y., Kildyakov A., Seil T., Finogeev E., Grabovoy A.* HWR200: New open access dataset of handwritten texts images in Russian. Computational Linguistics and Intellectual Technologies, 2023. Papers from the Annual International Conference "Dialogue". 2023. Issue 22. P. 452–458. https://doi.org/10.28995/2075-7182-2023-22-452-458
13. *Broder A.Z., Glassman S.C., Manasse M.S., Zweig G.* Syntactic clustering of the web. Computer networks and ISDN systems. 1997. V. 29 (8-13). P. 1157–1166.
14. *Broder A.Z.* On the resemblance and containment of documents. Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171). 1997. P. 21–29.
15. *Manber U., Myers G.* Suffix arrays: a new method for on-line search, SIAM Journal on Computing. 2003. V. 22.
16. *Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E.* Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research. 2011. V. 12. P. 2825–2830.
17. *Marti U.-V., Bunke H.* The IAM-database: an English sentence database for offline handwriting recognition. International Journal on Document Analysis and Recognition. 2002. V. 5. P. 39–46.
18. *Gatos B., Louloudis G., Causer T., Grint K., Romero V., Sánchez J.A., Toselli A.H., Vidal E.* Ground-Truth Production in the Transcriptorium Project, 11th IAPR International Workshop on Document Analysis Systems. 2014. P. 237–241.
19. *Toselli A.H., Romero V., Villegas M., Vidal E., Sánchez J.A.* HTR Dataset ICFHR. 2016. https://doi.org/10.5281/zenodo.1297399
20. *Potanin M., Dimitrov D., Shonenkov A., Bataev V., Karachev D., Novopoltsev M., Chertok A.* Digital Peter: New dataset, competition and handwriting recognition methods. The 6th International Workshop on Historical Document Imaging and Processing. 2021. P. 43–48.
21. *School_notebooks (2021)* Available at: https://github.com/ai-forever/htr_datasets/tree/main/school_notebooks.
22. *IDP-forms (*2021*)* Available at: https://github.com/ai-forever/htr_datasets/tree/main/IDP-forms.
23. *Toiganbayeva N., Kasem M., Abdimanap G., Bostanbekov K., Abdallah A., Alimova A., Nurseitov D.* KOHTD: Kazakh offline handwritten text dataset. Signal Processing: Image Communication. 2022. V. 108. P. 116827.

TEXT REUSE DETECTION IN HANDWRITTEN DOCUMENTS

A. Grabovoy[a,b], M. Kaprielova[a,b,c], A. Kildyakov[a], I. Potyashin[a], T. Seyil[a], E. Finogeev[a], and Y. Chekhovich[a,c]

[a] *Antiplagiat Company, Moscow, Russian Federation*

[b] *Moscow Institute of Physics and Technology, Moscow, Russian Federation*

[c] *FRC CSC RAS, Moscow, Russian Federation*

Presented by Academician of the RAS A.L. Semenov

Plagiarism detection in scholar assignments becomes more and more relevant nowadays. Rapidly growing popularity of online education, active expansion of online educational platforms for secondary and high school education create demand for development of an automatic reuse detection system for handwritten assignments. The existing approaches to this problem are not usable for searching for potential sources of reuse on large collections, which significantly limits their applicability. Moreover, real-life data is likely to be low-quality photographs taken with mobile devices. We propose an approach that allows to detect text reuse in handwritten documents. Each document is a picture and the search is performed on a large collection of potential sources. The proposed method consists of three stages: handwritten text recognition, candidate search and precise source retrieval. We represent experimental results for the quality and latency estimation of our system. The recall reaches 83.3% in case of better quality pictures and 77.4% in case of pictures of lower quality. The average search time is 3.2 seconds per document on CPU. The results show, that the created system is scalable and can be used in production, where fast reuse detection for hundreds of thousands of scholar assignments on large collection of potential reuse sources is needed. All the experiments were held on HWR200 public dataset.

Keywords: optical character recognition, handwriting, text reuse detection, computer vision, handwritten text recognition, plagiarism detection