АҚПАРАТТЫҚ-КОММУНИКАЦИЯЛЫҚ ТЕХНОЛОГИЯЛАР
ИНФОРМАЦИОННО-КОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ
INFORMATION AND COMMUNICATION TECHNOLOGIES

ИНЖЕНЕРЛІК БІЛІМ БЕРУДЕГІ ЦИФРЛЫҚ ТЕХНОЛОГИЯЛАР
ЖӘНЕ ЖАСАНДЫ ИНТЕЛЛЕКТ
ЦИФРОВЫЕ ТЕХНОЛОГИИ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ
В ИНЖЕНЕРНОМ ОБРАЗОВАНИИ
DIGITAL TECHNOLOGIES AND ARTIFICIAL INTELLIGENCE IN ENGINEERING EDUCATION

DOI 10.51885/1561-4212_2025_1_210
МРНТИ 28.23.15

Б.Ж. Сапуанов[1], Н.Ф. Денисова[2]
Восточно-Казахстанский технический университет имени Д. Серикбаева,
г. Усть-Каменогорск, Казахстан
[1]*E-mail: birzhan.sapuanov@gmail.com**
[2]*E-mail: NDenisova@ektu.kz*

АВТОМАТИЧЕСКИЙ ПОИСК МУЛЬТИЯЗЫЧНЫХ ДОКУМЕНТОВ АТТЕСТУЮЩИХСЯ УЧИТЕЛЕЙ ПОСРЕДСТВОМ ПОЛЕЙ ДАТЫ

МҰҒАЛІМДЕРДІҢ АТТЕСТАЦИЯДАН ӨТЕТІН КӨП ТІЛДІ ҚҰЖАТТАРЫН АВТОМАТТЫ ІЗДЕУ ҮШІН КҮН ӨРІСІН ҚОЛДАНУ

RETRIEVING DATE FIELDS IN MULTIPLE LANGUAGES FOR AUTOMATIC SEARCHING OF CERTIFIED TEACHER DOCUMENTS

Аннотация. *Различные школы проводят процессы сертификации учителей для продвижения, повышения и обновления их квалификации в соответствии с собственными правилами и процедурами. Данная процедура в Назарбаев Интеллектуальных школах (НИШ) Казахстана соблюдается без исключения. Важнейшим шагом в этом процессе является составление портфолио учителей, в котором содержатся многочисленные отсканированные документы, служащие доказательствами. В результате срок действия документов, предоставленных преподавателем, может быть просрочен, и они больше недействительны в течение периода сертификации. Внедрение системы, основанной на методах распознавания текста, имеет жизненно важное значение для ускорения процесса проверки документов учителей. В статье описаны методы, такие как глубокое обучение, искусственный интеллект и другие новые методы для улучшения процессов компьютерного зрения и распознавания текста, что, несомненно, повысило эффективность, инновационность и практичность процесса проверки.*

Каждый документ содержит ключевую информацию, такую как конкретное название сертификата, фамилия и имя. Цель данной статьи — представить систему автоматического извлечения полей даты из многоязычных письменных документов (казахский, английский и русский). Дата является одной из наиболее важных в информации, которую можно использовать во многих автоматизированных приложениях для индексации/поиска документов на основе даты. Чтобы разработать эту систему, сначала был определен сценарий, который находится в документе, и для каждой строки текста, относящейся к определенному сценарию, мы классифицируем словесные единицы по месяцам, а также с немесячными классами, применяя характеристики уровня слова, извлечение и классификацию. Затем выполняется сегментация немесячных слов на отдельные компоненты с последующей их маркировкой в виде цифр, текста, сокращений или знаков препинания. После этого в помеченных компонентах производится поиск возможных шаблонов дат, имеющихся в них. Для извлечения части даты использовались регулярные выражения как с числовыми, так и получисловыми частями. Классификация слов по месяцам и не по месяцам выполняется с использованием динамического искажения времени (DTW), а также подходов, основанных на признаках профиля. Цифры и знаки препинания в конечном итоге

обнаруживаются с учетом подхода на основе градиентных характеристик и классификатора машины опорных векторов (SVM). Эксперименты с наборами данных документов на казахском, английском и русском языках показали многообещающие результаты, полученные от предлагаемого подхода, что указывает на его эффективность.

Ключевые слова: *индексирование на основе даты, академические сертификаты, определение даты, извлечение даты, процедура аттестации учителей, многоязычные документы, искусственный интеллект.*

Аңдатпа. *Әртүрлі мектептер өздерінің ережелері мен рәсімдеріне сәйкес олардың біліктілігін көтеру, күшейту және жаңарту үшін мұғалімдерді аттестаттау процестерін жүргізеді. Қазақстандағы Назарбаев Зияткерлік мектептерінде (НЗМ) бұл процедура міндетті орындалады. Бұл процестегі ең маңызды қадам – мұғалімдердің портфолиосы, онда дәлел ретінде болатын көптеген сканерленген құжаттар бар. Нәтижесінде мұғалім ұсынған құжаттардың жарамдылық мерзімі өтіп, аттестаттау кезеңінде жарамсыз болып қалуы мүмкін. Тексеру процесінде мұғалімдердің құжаттарын жеделдету үшін мәтінді тану әдістеріне негізделген жүйені енгізу өте маңызды. Бұл мақалада терең оқыту, жасанды интеллект және басқа да жаңа әдістер компьютерлік көру мен мәтінді тану процестерін жақсарту үшін қолданылды, бұл тексеру процесінің тиімділігін, инновациясын және қолданылуын арттырды.*

Әрбір құжатта сертификаттың нақты атауы, тегі және аты сияқты негізгі ақпарат бар. Бұл мақаланың мақсаты – көптілді жазба құжаттардан (қазақ, ағылшын және орыс) күн өрістерін автоматты түрде алу жүйесін ұсыну. Күн көптеген автоматтандырылған қолданбаларда уақыт негізінде құжаттарды индекстеу/іздеу ретінде пайдаланылуы мүмкін ең маңызды ақпараттың бірі болып табылады. Бұл жүйені жобалау үшін алдымен оның астында құжат келетін сценарий анықталды және анықталған сценарийге қатысты мәтіннің әрбір жолы үшін сөз бірліктерін айға, сонымен қатар сөз деңгейіндегі сипаттаманы қолданатын айға сәйкес емес сыныптарға жіктейміз. Айға сәйкес емес сөздерді жеке құрамдас бөліктерге бөлу, содан кейін олардың сан, мәтін, қысқарту немесе тыныс белгілері ретінде белгіленуі орындалады. Осыдан кейін тегтелген құрамдас бөліктер олардың ішінде қол жетімді ықтимал күн үлгілерін іздейді. Сандық және жартылай сандық бөліктері бар тұрақты өрнектердің екеуі де күн бөлігін шығару үшін пайдаланылды. Айлық және айлық емес сөздерді жіктеу динамикалық уақытты өзгерту (DTW), сондай-ақ профиль мүмкіндіктеріне негізделген тәсілдерді пайдалану арқылы орындалады. Цифрлар мен тыныс белгілері градиент негізіндегі сипаттамалық тәсілге және қолдау векторлық машинасының (SVM) классификаторына қатысты анықталады. Қазақ, ағылшын және орыс құжаттарының деректер жинақтары бойынша жүргізілген тәжірибелер ұсынылған тәсілдің тиімділігін көрсететін перспективалы нәтижелерді көрсетті.

Түйін сөздер: *күнге негізделген индекстеу, академиялық сертификаттар, күнді анықтау, күнді алу, мұғалімдерді аттестациялау процедурасы, көп тілді құжаттар, жасанды интеллект.*

Abstract. *Different schools conduct teacher certification processes to promote, reinforce, and innovate their qualification in accordance with own rules and procedures. This procedure at the Nazarbayev Intellectual schools (NIS) in Kazakhstan is followed without exception. The most important step in this process is the compilation portfolio of the teachers, where it contains numerous scanned documents that serve as evidences. As a result, teacher's provided documents may have expired and are no longer valid during the certification period. The implementation of a system ground on text recognition techniques is vital to accelerate teachers' documents in verification process. In this article, deep learning, artificial intelligence and new other methods have been applied to improve computer vision and text recognition processes, which undoubtedly have increased the efficiency, innovation, and practicality of the verification process.*

Each document has key information, such as particular name of certificate, family name and first name. The purpose of this article is to present a system for automatic extraction of the date fields from the multilingual written documents (Kazakh, English and Russian). The date is one of the most important information, which can be used in many automated applications, as document indexing/retrieval on basis of date. To design this system, first it was identified the script, which a document comes below it, and for each line of text that relates to an identified script, we classify word units into month, also with non-month classes applying word-level characteristic extraction and classification. Segmentation of non-monthly words into individual components followed by their labeling as digit, text, contraction or punctuation is then done. After that, tagged components are searched for possible date patterns available within them. Both regular expressions with numeric as well as semi-numeric parts have been used to extract date part. Month and non-month words classification is performed by utilizing Dynamic Time Warping (DTW) as well as profile feature-based approaches. Digits and punctuation marks are detected with respect to gradient-based characteristic approach and Support Vector Machine (SVM) classifier eventually. Experiments on Kazakh, English and Russian document datasets have shown promising results obtained from the proposed

approach indicating its effectiveness.

Keywords: *date-based indexing, academic certificates, date spotting, date extraction, teacher's certification procedure, multi-lingual documents, artificial intelligence.*

Введение. Процедура аттестации учителей школ НИШ состоит из трех основных этапов, каждый из которых требует многочисленных доказательств, содержащих тексты, изображения или и то, и другое. Также существует возможность того, что каждый документ может быть на трех разных языках (казахском, английском или русском), поскольку в школе практикуется трехъязычная политика. Таким образом, все эти факторы затрудняют проверку портфолио учителей заместителем директора. В частности, когда кандидаты предоставляют документы с истекающим сроком действия. Преподаватель не может быть допущен к аттестации из-за технических или человеческих ошибок.

В настоящее время систематизация имеющихся в организации управленческих документов по датам требует слишком большой кропотливой работы и отнимает много времени. Например, один заместитель директора школы НИШ, отвечающий за продвижение учителей (всего их 22 по всей сети НИШ), проверяет каждый год не менее 30-35 портфолио, содержащих около 80-100 страниц на 3-х языках. Поэтому необходимо рассмотреть возможность применения автоматизации при решении вышесказанных рутинных задач.

В данной статье предлагается использовать автоматизированный подход на основе дат для индексации соответствующих документов: представлена автоматическая система получения многоязычных (на казахском, английском и русском языках) полей дат документа. Дата является важным разделом информации и может использоваться в качестве ключа для различных приложений, включая документы, основанные на датах для поиска или индексации хранилищ записей, в качестве административных документов, дипломов и сертификатов. Кроме того, автоматическое извлечение данных о датах затруднено, поскольку каждый сертификат формируется различными организациями, имеются неоднозначные символы между цифрами и буквами, а также знаки препинания, которые могут привести к неправильной классификации при цифровой идентификации. В многоязычных школах, таких как NIS, извлечение многострочных документов с использованием шаблона даты может оказаться важным. Поэтому в одном документе может быть один или несколько сценариев, которые на 3-х языках. На рис. 1 показаны различные сертификаты и свидетельства учителей. Поскольку наш метод работает на нескольких языках, он содержит четыре основные части: детекторы слов месяцев, поля для чисел, идентификацию алфавита и извлечение шаблонов дат. В каждом сертификате учитываются два типа шаблонов для поля даты (числовой и буквенно-цифровой). Поэтому предлагаемый процесс извлечения данных из таких документов будет очень полезен в их поиске и понимании.

При обработке многоязычных документов первым делом необходимо определить, какой языковой сценарий используется, идентифицировать слова по месяцам и найти цифры/знаки препинания, которые включают косую черту («/»), дефис («-») или точку. («.»). Недавно было выполнено несколько работ над многоязычными и многошрифтовыми документами для извлечения даты из полей. Однако, чтобы получить представление о том, что происходит при идентификации шрифтов и определении слов с целью извлечения числовых полей и распознавания дат, здесь будут представлены старые методы, а также несколько недавно разработанных подходов.

Идентификация сценария является полной для исследовательских работ, особенно в печатных и рукописных (офлайн/онлайн) документах. Предыдущая работа по идентификации скриптов разделяла их на три типа категорий: уровни блока/абзаца, строки и слова (Obaidullah et al., 2019). Согласно (Spitz et al., 1997), существует двухэтапный метод

автоматической идентификации шрифтов для цифровых отпечатков, а также языка, на котором они написаны. Первым шагом является классификация шрифтов на две общие группы (ханьские или латинские) с использованием распределения вертикального положения информации о вогнутости вверх. Более того, по распределению информации об оптической плотности можно узнать, какой из корейских, японских и китайских языков принадлежит к классу ханьской письменности. С другой стороны, языки на основе латиницы предполагают использование кодов формы символов для их идентификации (Mandal et al., 2015). Для распознавания письменности и языка, используемых в рукописных документах, (Ferrer et al., 2024) предложили подход, основанный на связанных компонентах. Каждый компонент дает несколько эвристических функций, которые извлекаются с помощью линейного дискриминантного анализа (LDA), используемого для классификации каждой возможной пары алфавитов в наборе данных (арабский против китайского, арабский против кириллицы и т.д.) (Gambella et al., 2021).

Рисунок 1. Примеры удостоверений учителей (фактические даты отмечены красными/синими квадратиками, и черные ящики скрывают личные данные учителей)
Примечание – составлено авторами

Среди процедур, предлагаемых для разделения рукописных документов, есть предложения (Liang et al., 2019), основанные на принципах геометрической структуры, зон занятости и топологии. В этой работе для классификации использовалась нейронная сеть (НС). (Roy et al., 2010) предложили метод идентификации почерка латинского алфавита с почерком персидских букв. Было использовано около 12 характеристик, таких как

фрактальные размерности или положение мелких компонентов, или даже связность.

Обнаружение слов – важная область исследований, в которой была проделана большая работа по обеспечению возможности поиска и просмотра рукописных документов (Sushma et al., 2024; Omayio et al., 2023). По этой причине определение слов набирает популярность из-за низкой стоимости вычислений по сравнению с расшифровкой целых текстов. Методы определения слов, основанные на функциях на основе профиля, а также динамическое искажение времени (DTW), были предложены Шриваставой и Харитом в их статье (Srivastava et al., 2020). Кроме того, рукописные документы могут быть доступны для поиска и индексирования с помощью подхода, основанного на рекуррентной нейронной сети (RNN) (Ghosh et al., 2019). Нейронные сети (NN), алгоритмы коннекционистской временной классификации (CTC) и передачи токенов были инструментами, используемыми для определения слов (Scheidl et al., 2018).

Метод скрытой марковской модели (HMM) часто используется при моделировании рукописного текста, обнаружении слов и в других случаях. (Fischer et al., 2013) представили подход к обнаружению слов, основанный на обучении, который использует модели подслов HMM для определения ключевых слов. Предлагаемый здесь метод не требует словарного запаса и позволяет обнаружить любые ключевые слова в рукописи. Идея определения слов в рукописных документах с использованием HMM была реализована в (Rothacker et al., 2013). В этой работе применялись функции локальной градиентной гистограммы (LGH).

Предлагаемая работа продвигается дальше в объяснении документов и использует метки распознавания символов в казахских, английских и русских буквенно-цифровых словах для установления связи с полями даты в многоязычных документах. Недавно Мандал и др. предоставили подход для извлечения поля даты из рукописных английских документов (Mandal et al., 2012). Настоящая работа представляет собой расширение, касающееся извлечения данных из многоязычных рукописных документов учителя на трех языках. Насколько нам известно, существует несколько работ по извлечению даты в печатных/рукописных документах на разных языках. Обнаружение и интерпретация образца даты в рукописных записях затруднены из-за разных стилей письма в разных организациях, а также образцов одной даты, и т.д.

Материалы и методы исследования. Здесь предлагается многоэтапный подход к извлечению данных из месторасположения записей в документах. Вначале сценарии обнаруживаются с использованием переднего плана и деталей фона. Для выделения переднего и заднего планов эта система использует метод, основанный по принципу резервуара. Верхний и нижний резервуары слов используются для разделения слов на примитивные сегменты, и эти сегменты классифицируются с помощью SVM как примитивные сегменты английского, казахского или русского языков. Сценарий идентифицируется на основе большинства классифицированных примитивных сегментов. После этого система обучается относительно своей модели после идентификации сценария (Littell et al., 2019). Каждый сценарий имеет две модели (месяц и цифра), которые обучают систему двум этапам классификации. На следующем этапе рукописные фрагменты слов месяца/не месяца различаются друг от друга. Для этой цели блоки слов извлекались с помощью морфологических операций, в то время как другие подвергались анализу характеристик степени блока слов, прежде чем их классифицировали как категории месяцев или не месяцев (Ravi et al., 2013). В-третьих, для немесячных блоков слов рукописи проводится анализ части блока слов. Анализ и распознавание характеристик на уровне компонентов указывают на отдельные алфавиты, цифры и знаки препинания. Однако те компоненты, распознавание которых имеет низкую достоверность, необходимо дополнительно проанализировать на предмет классификации разделов строк записей (Roy

et al., 2012). Для установления характеристик использовался 400-мерный градиент наряду с методом классификации блоков на основе DTW. Кроме того, здесь были введены SVM и в классификацию на уровне разделов (Mandal et al., 2012). Наконец на последнем этапе из последовательности помеченных компонентов наблюдаются числовые и получисловые (содержат поле месяца в виде текстовой строки) образцы дат. Следовательно, для выбора строк-кандидатов сначала использовался метод голосования, за которым следовал анализ регулярных выражений, который выявлял закономерности дат.

Применение методов DTW и SVM было реализовано с помощью языка программирования Python 3.8 в среде IDE PyCharm Community Edition с использованием библиотек Tesseract OCR, math, recognizers-text-suite, EasyOCR, Pillow, OpenCV + Keras.

Извлечение поля даты. Здесь обсуждался подход, основанный на обнаружении, для поиска поля даты в документе. Поиск даты происходит на основе нахождения строк, содержащих даты. Во-первых, идентифицируются различные части, составляющие дату, такие как название месяца, цифры и знаки препинания, и они используются для извлечения возможных дат. Выполняется поиск по различным формам календарных дат, и в этом разделе описывается, как они были обнаружены, включая другие составляющие и целые поля, из которых они состоят.

Компоненты даты. Компонентами поля даты в документе могут быть цифры, текстовый месяц, знаки препинания и сокращения. Шаблоны дат могут иметь две категории на основе разных типов компонентов (числовые и получисловые даты). Числовая дата: поля числовой даты состоят из цифр и знаков препинания (примеры числовых дат: 08/06/2024, 8/6/2024, 8-6-24 на английском языке; 08.06.24 или 08.06.2024 на казахском языке; 06.08.24 или 06.08.2024 на русском языке). Из набора данных было замечено, что общее количество компонентов в поле для числовых дат может варьироваться от 6 до (дата записывается как 06.08.24, где она содержит 6 компонентов, или как 08/06/2024, где компонентов будет десять). Регулярное выражение предстоящей даты демонстрирует допустимые форматы числовой даты:

$$(\mathrm{d}|\mathrm{dd})(/.-)(\mathrm{d}|\mathrm{dd})(/.-)(\mathrm{dd}|\mathrm{dddd}),$$

здесь d означает цифру. 3 части или поля имеют дату, например, поле даты, месяца и года. Правильное поле числовой даты содержит одно - или двузначную запись дня, одно- или двузначную запись месяца, а также двузначную или четырехзначную запись года. Кроме того, в числовом типе даты должны быть два знака препинания, разделяющие запись года, месяца и дня.

Классификация слов месяца. В данной работе для определения месячных/немесячных терминов используется модель на основе DTW (Mandal et al., 2015). Сходство между двухмесячными последовательностями измеряется с использованием метода DTW. Это означает, что с помощью этого метода оцениваются сходства между двухмесячными последовательностями. Фактически он измеряет степень сходства любой пары месяцев. В результате можно измерить, насколько близко совмещены или различны 2 последовательности, находящиеся во времени, принимая во внимание определенные временные искажения, включая растяжение или сжатие. Речь, подписи и даты, написанные в виде текста, входят в число многих других приложений, где этот метод широко используется. Другими словами, слово может быть выражено в виде четырех последовательностей, которые можно описать иначе, чем указано выше. Полоса (Sakoe et al., 1978) помогает ускорить вычисления для измерения сходства между двумя рассматриваемыми последовательностями с использованием этого подхода. Например, если слово-кандидат имеет ширину 4 пикселя, то оно не будет рассматриваться, если его изображение шире 8 пикселей.

Указанный метод относится к системе, которая применяет DTW к 2 сигналам, содержащим 4 последовательности признаков $(f_k, where\ 1 \le k \le 4)$, которые представляют собой признак верхнего профиля, признак нижнего профиля, профиль вертикальной проекции и переход слов от фона к переднему плану. Матрица D используется для расчета расстояния DTW между $I1$ и $I2$ сигналы (Mandal et al., 2015):

$$D(i,j) = \min\begin{pmatrix} D(i,j-1) \\ D(i-1,j) \\ D(i-1,j-1) \end{pmatrix} + d(x_i, y_i)$$

$$d(x_i, y_i) = \sum_{k=1}^{4}\left(f_k(I_1, i) - f_k(I_2, j)\right)^2 \quad (1)$$

Рассмотрим расстояния совпадений для четырех характеристик слова. Просуммировав их, можно получить совокупное расстояние, которое является окончательной стоимостью совпадения слова. Наконец, эта стоимость сопоставления приравнивается к длине пути деформации. Для определения месячных полей в этом документе использовались расстояние DTW и метод классификации на основе ближайшего соседа. На рис. 2 текстовые строки представлены в виде групп месяцев, а также блоков, не относящихся к месяцам.

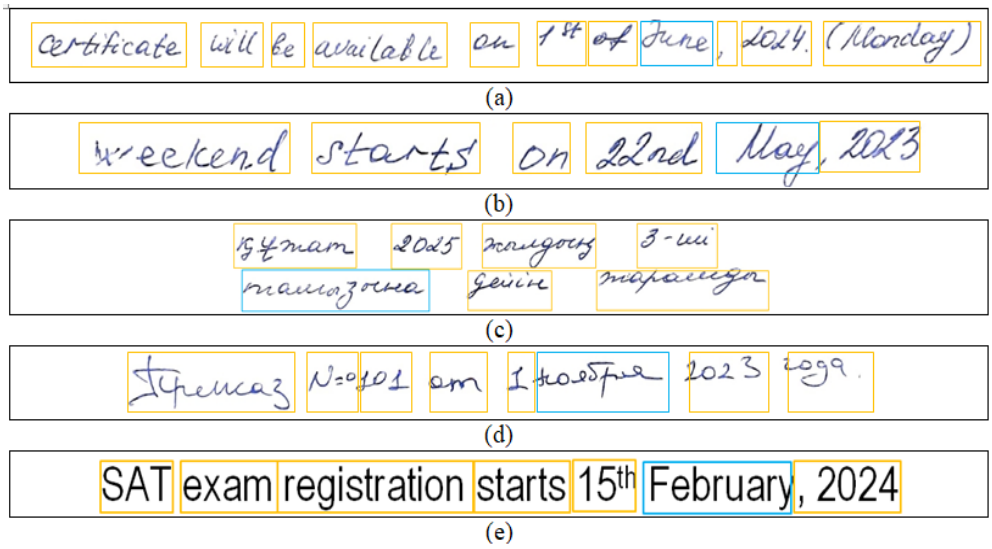

Рисунок 2. Рукописные строки текста на английском языке демонстрируют идентификацию блоков месяцев (a и b). Казахские рукописные тексты (c). Русские рукописные тексты (d). Английские печатные тексты (e). Месячные и немесячные блоки отмечены оранжевым и синим цветами соответственно

Примечание – составлено авторами

Выбор линии кандидата. Строки текста, в которых указаны месяцы, имеют цифры и знаки препинания. В этой строке подсчитывалось целое количество цифр, строк месяца и знаков препинания в текстовой строке. По подсчетам по каждому предмету были выявлены закономерности даты из этой строки. Допустимые шаблоны дат для созданного экспериментального набора данных содержат как минимум шесть элементов, показанных в примерах (например, на казахском языке: 08 маусым 2022 жыл. На русском языке: 08 июня 2022 года, на английском языке: Jun 8th, 22). Если в строке содержится шесть или более элементов даты, она считается одной из строк-кандидатов, которые будут использоваться для поиска шаблона даты.

Результаты и их обсуждения. Насколько было установлено, не существует стандартного набора данных рукописных дат для оценки методов извлечения дат. Поэтому при обучении классификаторов на разных уровнях используются различные наборы данных. В табл. 1 представлена сводка данных обучения и тестирования, которые использовались в экспериментах. Важно понимать, что наши экспериментальные обучающие и тестовые наборы данных не совпадают.

Набор обучающих данных: в классификаторе SVM уровня идентификации алфавита для идентификации казахского, английского и русского алфавита распознано не менее 35 шаблонов для каждого класса цифр для наборов обучающих данных казахских, английских и русских чисел. Этот набор данных был собран ранее в школе для систем распознавания рукописного текста. Из этих рукописных слов в ходе тренингов генерировались примитивные сегменты, которые обучали классификатор распознавать сценарии. Например, данные за месяц имеют различные форматы (к примеру, ОКТЯБРЬ, 10, окт. и X) и были собраны из 35 документов. Напомним, что эти 48 классов английского месяца записаны буквами алфавита (OCTOBER, 10, OCT. и X), а 24 класса для казахского и русского алфавитов рассматриваются в наших экспериментах как данные для классификации месяцев с помощью классификатора, полученного в ходе обучения, а это в свою очередь еще больше затрудняет разпознование записей на разных языках.

Следующие символы, такие как цифры и знаки препинания, включены для обучения классификаторов идентификации уровня компонентов определению дат. Эти числа включают строки с 5 по 7 в табл. 1, где они указывают количество обучающих выборок, использованных для разработки системы распознавания уровней символов. Классификатор уровня компонента был обучен с использованием английских чисел из базы данных рукописных цифр MNIST (Roy et al., 2010) во всех алфавитах. Опять же, помните, что, как уже говорилось ранее, иногда в казахских или русских датах используются английские цифры, поэтому использовались как цифры MNIST вместе с цифрами собственного алфавита, так и при обучении всем трем символам. Кроме того, рассматривались наборы данных для обучения казахским или русским цифрам, а также английские цифры MNIST для целей обучения как казахскому, так и русскому алфавиту. Также были созданы числовые данные для использования с казахской и русской письменностью, которые были разработаны ранее в школе для использования во время обучения. Здесь для каждого класса чисел было распознано не менее 35 шаблонов для наборов данных для обучения цифрам на казахском, английском и русском языках.

Таблица 1. Набор экспериментальных данных

Типы данных	Казахский	Русский	Английский
Данные обучения			
Слово/Текст	1127	1041	1278
Месяц	912	972	1032
Цифры/цифры	2210	2103	Набор данных MNIST
Пунктуация	831	714	703
Сокращение	458	536	895
Данные испытаний			
Рукописная линия	1201	1125	1371
Примечание – составлено авторами			

Этот набор данных был протестирован с использованием строк рукописи из одного сценария и информации из нескольких сценариев, собранной от 35 учителей. Общее количество рукописных строк, собранных из этих трех сценариев для объемов

тестирования, указано в строке 9 табл. 1.

Выходные данные идентификации сценария. Эксперименты по идентификации сценария были разработаны для выбора функций. Вычисленные результаты идентификации сценария с применением двух функций в экспериментальном наборе данных показаны в табл. 2. Обоснованная SVM ($\gamma = 1.5$) пятикратная перекрестная проверка достоверности приведена для градиентных функций и функций на основе фильтра Габора. Например, он вычисляет точно такой же признак на основе фильтра Габора, упомянутый в этой статье [19]. В табл. 2 показано, как примитивные сегменты переднего плана (FS) и фоновые капли-резервуары (RB) по отдельности способствуют решению задачи идентификации сценария.

Таблица 2. Результат идентификации сценария на основе SVM, точность теста основана на резервуарных объектах и примитивных сегментах (пятикратная перекрестная проверка)

Функция на основе	Примитивные сегменты (%)	Капли резервуара (%)
Градиент	71,23	58,52
Габор	62,74	49,95
Примечание – составлено авторами		

Более того, здесь также можно увидеть достижение обеих характеристик в этом эксперименте. Эти результаты побудили нас использовать функции градиента вместо функций Габора в случае идентификации скрипта, поскольку градиент превзошел Габора во время пятикратных перекрестных проверочных тестов на основе SVM. Градиент работает лучше, чем Габор, для идентификации сценария, поэтому вместо него использовался градиент Габора, который обеспечивал максимальную точность 62,74 %, когда в качестве входных данных использовалась информация примитивных сегментов, и минимальную точность 71,23 % при тех же условиях, но для функции градиента.

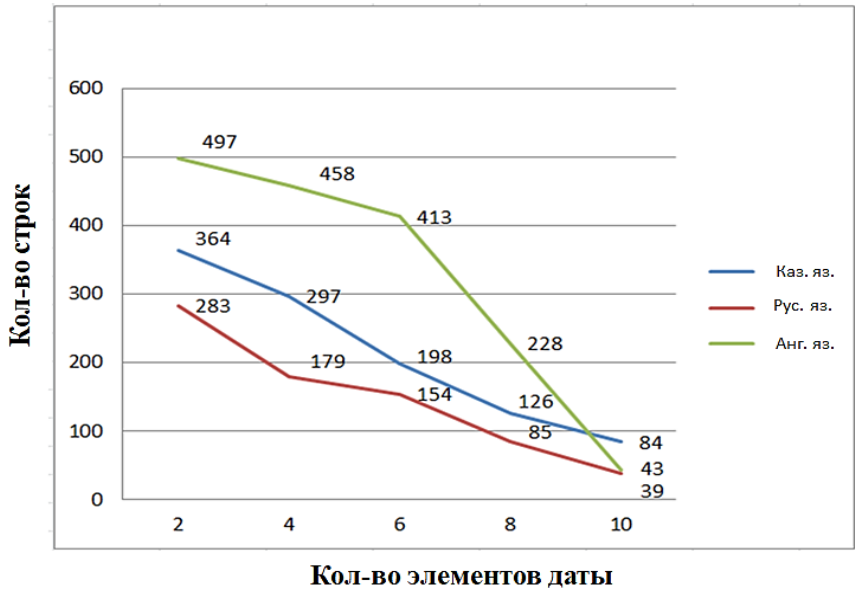

Рисунок 3. На графике представлены результаты строчных фильтров на казахском, английском и русском алфавитах. Он проверяет строки на наличие от 2 до 10 элементов.

Примечание – составлено авторами

Поле извлечения выходных данных даты. Определение количественной производительности нашей системы при извлечении дат из трех упомянутых выше сценариев происходит на основе Точности (P) и полноты (R) извлечения. Для более точной и полной степени распознования строки, содержащей даты, требуются входные данные в виде элементов даты и строк (рис. 3). Они требуются, чтобы оценить точность и степень полноты извлекаемых данных из полей, содержащих даты.

Заключение. Систему для автоматической проверки документов аттестуемых учителей можно обучить автоматическому поиску на основе собранной информации о датах, используя такой подход, как представленный здесь, с использованием машинного интеллекта или распознавания образов при извлечении дат из рукописных записей, состоящих из нескольких рукописных символов. Используя метод на основе DTW, в данной статье представлен способ извлечения текстовых элементов месяца из рукописей сценариев. Атрибуты на основе градиента вместе с SVM используются для распознавания различных компонентов даты, таких как знаки препинания, цифры и сокращения. В конце строка помеченных компонентов ищет различные шаблоны дат. В целом этот метод дает положительные результаты, особенно с учетом того, что это первоначальная работа над рукописным многоязычным документом, в котором поля даты извлекаются автоматически. В будущих работах было бы полезно уменьшить ошибки сегментации на уровне слов и символов.

Чтобы получить лучшие результаты, подходы без сегментации на основе HMM могут исправить эти ошибки. Более того, такие методы могут улучшить уровень распознавания во всем. Использование единого метода извлечения признаков и классификатора поможет разработать лучшую систему с точки зрения простоты. Такие достижения расширят возможности поиска с помощью регулярных выражений. Таким образом, можно сделать вывод, что дальнейшие исследования могут быть направлены на достижение общих решений в анализе документов, а это означает, что применение автоматизации извлечения и распознования полей дат может найти новые применения.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

«Уведомление об использовании генеративного ИИ и технологиях с его помощью в процессе написания рукописи». При подготовке данной работы авторы не использовали генеративный ИИ.

Список литературы

Ferrer M.A., Das A., Diaz M., Morales A., Carmona-Duarte C., Pal U. (2024). MDIW-13: a New Multi-Lingual and Multi-Script Database and Benchmark for Script Identification. Cognitive Computation, vol. 16, 131-157, https://doi.org/10.1007/s12559-023-10193-w.

Fischer A., Frinken V., Bunke H., Suen C.Y. (2013). Improving hmm-based keyword spotting with character language models. 12th International Conference on Document Analysis and Recognition, 506-510, https://doi.org/10.1109/ICDAR.2013.107

Gambella C., Ghaddar B., Naoum-Sawaya J. (2021). Optimization problems for machine learning: A survey. European Journal of Operational Research, vol. 290, 807–828,
https://doi.org/10.1016/j.ejor.2020.08.045.

Ghosh R., Vamshi C., Kumar P. (2019). RNN based online handwritten word recognition in Devanagari and Bengali scripts using horizontal zoning. Pattern Recognition, vol. 92, 203–218, https://doi.org/10.1016/j.patcog.2019.03.030.

Liang T., Poggio T., Rakhlin A., Stokes J. (2019). Fisher-Rao metric, geometry, and complexity of neural networks. In Proceeding of the international conference on artificial intelligence and statistics, vol. 89, 888–896, https://doi.org/10.48550/arXiv.1711.01530.

Littell P., Lo Ch., Larkin S., Stewart D. (2019). Multi-Source Transformer for Kazakh-Russian-English Neural

Machine Translation. In Proceedings of the Fourth Conference on Machine Translation, vol. 2, 267–274, https://doi.org/10.18653/v1/W19-5326.

Mandal R., Roy P., Pal U. (2012). Date field extraction in handwritten documents. 21st International Conference on Pattern Recognition (ICPR), 533-536,doi:10.1109/ICDAR.2015.7333885.

Mandal R., Roy P., Pal U. (2012). Signature segmentation from machine printed documents using contextual information. International Journal of Pattern Recognition and Artificial Intelligence. – Vol. 26, https://doi.org/10.1142/S0218001412530035.

Mandal R., Roy P.P., Pal U., Blumenstein M. (2015). Multi-lingual date field extraction for automatic document retrieval by machine. Information Sciences. – Vol. 314, 277-292, https://doi.org/10.1016/j.ins.2014.08.037.

Obaidullah S., Santosh K., Halder C., Das N., Roy K. (2019). Automatic Indic script identification from handwritten documents: page, block, line and word-level approach. Int. J. Mach Learn Cybern, vol. 10, 87-106, https://doi.org/10.1007/s13042-017-0702-8.

Omayio E., Indu S., Panda J. (2023). Word spotting and character recognition of handwritten Hindi scripts by Integral Histogram of Oriented Displacement (IHOD) descriptor. Multimedia Tools and Applications, 1-30, https://doi.org/10.1007/s11042-023-15219-x.

Ravi S., Khan A. M. (2013). Morphological Operations for Image Processing: Understanding and its Applications. Proc. 2nd National Conference on VLSI, Signal processing & Communications NCVSComs-2013.

Rothacker L., Rusinol M., Fink G.A.(2013). Bag-of-features HMMs for segmentation-free word spotting in handwritten documents. 12th International Conference on Document Analysis and Recognition, 1305-1309, https://doi.org/10.1109/ICDAR.2013.264.

Roy K., Alaei A., Pal U. (2010). Word-wise handwritten Persian and Roman script identification. In Proceedings of the International Conference on Frontiers in Handwriting Recognition (ICFHR), 628–633, https://doi.org/10.1109/ICFHR.2010.103.

Roy P., Pal U., Lladós J., Delalandre M. (2012). Multi-Oriented Touching Text Character Segmentation in
Graphical Documents using Dynamic Programming. Pattern Recognition, vol. 45, 1972-1983, https://doi.org/10.1016/j.patcog.2011.09.026.

Sakoe H., Chiba S. (1978). Dynamic Programming Algorithm Optimization for Spoken Word Recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 26, 43-49, https://doi.org/10.1109/TASSP.1978.1163055

Scheidl H., Fiel S., Sablatnig R.(2018). Word Beam Search: A Connectionist Temporal Classification Decoding Algorithm. 16th International Conference on Frontiers in Handwriting Recognition (ICFHR), 253-258, doi: 10.1109/ICFHR-2018.2018.00052.

Spitz A.L. (1997). Determination of script, language content of document images. Patt. Recog., vol. 19, 235–245, doi: 10.1109/34.584100.

Srivastava D., Harit G. (2020). Word Spotting in Cluttered Environment. Advances in Intelligent Systems and Computing, vol. 1024, 161–172, https://doi.org/10.1007/978-981-32-9291-8_14.

Sushma S., Sharada B. (2024). Two-Stage Word Spotting Scheme for Historical Handwritten Devanagari Documents. In Data Analytics and Learning, vol. 779, 1–18, https://doi.org/10.1007/978-981-99-6346- 1_1.

Information about authors

Sapuanov Birzhan – 2nd year doctoral student in Information systems, D. Serikbayev East Kazakhstan technical university, Oskemen, Kazakhstan, E-mail: birzhan.sapuanov@gmail.com, +7 778 690 73 25

Natalia Denissova – Candidate of Physical and Mathematical Sciences, Associate Professor, East Kazakhstan Technical University, Oskemen, Kazakhstan, E-mail: NDenisova@ektu.kz