



Методология исследования геосистем Methodology of Geosystems Research

УДК 004.93'1:912.43:81'373.21

DOI 10.52575/2712-7443-2025-49-3-477-499

EDN KKFUTD

Автоматическое распознавание топонимов: новые вызовы для цифровой картографии

Крупнова Е.С., Дмитриев А.В.

Санкт-Петербургский политехнический университет Петра Великого
Россия, 195251, г. Санкт-Петербург, ул. Политехническая, 29Б
krupnalena@mail.ru, avd84@list.ru

Аннотация. В статье представлен комплексный анализ современных методов автоматического распознавания топонимов на географических картах и в текстах. Проведено сравнение вычислительных подходов, включая графовые методы (MST), ансамблевые алгоритмы, морфологические операции и лексико-семантические методы. Авторы рассматривают вызовы, связанные с многословными топонимами, неоднозначностью географических названий и визуальными особенностями карт. Особое внимание уделяется метрикам оценки эффективности и вычислительной сложности алгоритмов. Проведена систематизация методологических этапов обработки топонимов с учетом российских и зарубежных исследований. Исследование выявляет ключевые проблемы в области распознавания географических названий: высокую плотность информации на картах, пересечение текстовых меток с другими картографическими элементами, разнообразие шрифтов, размеров и ориентаций текста. Анализируются вычислительные особенности и точность различных подходов, от традиционных методов до современных архитектур глубокого обучения. Предложена концепция комплексной архитектуры автоматического распознавания топонимов, интегрирующая методы компьютерного зрения, обработки естественного языка и пространственного анализа. Авторы демонстрируют, что наиболее перспективным направлением является создание адаптивных систем, способных динамически выбирать оптимальный набор алгоритмов в зависимости от характеристик обрабатываемых данных и задач конкретного исследования.

Ключевые слова: топонимы, распознавание текста, геоинформационные системы, исторические карты, компьютерное зрение

Для цитирования: Крупнова Е.С., Дмитриев А.В. 2025. Автоматическое распознавание топонимов: новые вызовы для цифровой картографии. Региональные геосистемы, 49(3): 477–499.
DOI: 10.52575/2712-7443-2025-49-3-477-499 EDN: KKFUTD

Automatic Recognition of Toponyms: New Challenges for Digital Cartography

Elena S. Krupnova, Alexander V. Dmitriev

Peter the Great St. Petersburg Polytechnic University
29B Polytechnicheskaya St, St. Petersburg 195251, Russia
krupnalena@mail.ru, avd84@list.ru

Abstract. The article presents a comprehensive analysis of modern methods for automatic toponym recognition on geographic maps and in texts. The authors compare computational approaches including



graph methods (MST), ensemble algorithms, morphological operations, and lexico-semantic methods. The paper examines challenges related to multi-word toponyms, geographical name ambiguity, and visual features of maps. Special attention is given to efficiency evaluation metrics and computational complexity of algorithms. The study systematizes methodological stages of toponym processing considering both Russian and international research. The research identifies key challenges in geographic name recognition: high information density on maps, intersection of text labels with other cartographic elements, and diversity of fonts, sizes, and text orientations. The paper examines computational characteristics and accuracy of various approaches, from traditional methods to modern deep learning architectures. The authors propose a concept of an integrated architecture for automatic toponym recognition that combines computer vision methods, natural language processing, and spatial analysis. The study demonstrates that the most promising direction is the development of adaptive systems capable of dynamically selecting the optimal set of algorithms depending on the characteristics of the processed data and the objectives of specific research.

Keywords: toponyms, text recognition, geographic information systems, historical maps, computer vision

For citation: Krupnova E.S., Dmitriev A.V. 2025. Automatic Recognition of Toponyms: New Challenges for Digital Cartography. *Regional Geosystems*, 49(3): 477–499 (in Russian). DOI: 10.52575/2712-7443-2025-49-3-477-499 EDN: KKFUTD

Введение

Карты, будучи визуальным отражением пространственных отношений и человеческих знаний о мире, представляют собой уникальную форму документирования историко-культурной и географической информации [Kent, 2008]. Исторические карты особенно ценны, поскольку запечатлевают не только геопространственную реальность прошлого, но и отражают лингвистические, политические и культурные трансформации через изменения топонимов [Olson et al., 2023].

В контексте современной цифровой гуманитаристики и геоинформатики обработка и анализ картографических материалов приобретают новое измерение благодаря передовым методам компьютерного зрения и машинного обучения [Hu et al., 2022a].

Автоматическое распознавание топонимов на картах представляет собой комплексную междисциплинарную задачу, находящуюся на пересечении компьютерного зрения, обработки естественного языка, исторической географии и геоинформационных систем [Chiang et al., 2014]. Хотя прогресс в области оптического распознавания символов (OCR) достиг значительных результатов, применение этих технологий к картографическим материалам остается нетривиальной задачей из-за уникальных особенностей карт как визуально-текстуальных документов [Bluche, 2015; Schlegel, 2021].

Общегеографические и исторические карты характеризуются высокой плотностью информации, где текстовые метки пересекаются с другими картографическими элементами, имеют разнообразные шрифты, размеры, ориентации и распределение по поверхности [Pezeshk, Tutwiler, 2011]. Как отмечают Deseilligny с соавторами [1995], текстовые метки на картах могут появляться в разных цветах, ориентациях, размерах и шрифтах, а неоднородный фон и перекрытие с другими элементами карты существенно снижают точность распознавания.

В отличие от стандартных задач OCR, распознавание топонимов на картах требует синтеза различных подходов, включая детекцию текста, распознавание символов, геокодирование и разрешение топонимической неоднозначности [Ardanuy, Sporleder, 2017]. Эта неоднозначность усугубляется тем, что многие топонимы являются полисемичными — одно название может относиться к нескольким географическим объектам. Kytamargiou с соавторами [2020] подчеркивают значимость изучения изменений топонимов для понима-

ния культурно-исторических процессов, что делает автоматическое распознавание особенно ценным инструментом для истории картографии и пространственной гуманитаристики.

Цель исследования – создание и обоснование комплексной методологии автоматического распознавания топонимов на картографических материалах через интеграцию методов компьютерного зрения, обработки естественного языка и пространственного анализа для повышения точности идентификации многословных и низкочастотных географических названий.

Объекты и методы исследования

Объектом исследования выступает процесс автоматического распознавания топонимов на картографических материалах и в текстах. Топонимы рассматриваются как лингвистические единицы с пространственными, историческими и культурными характеристиками. Особое внимание уделяется историческим картам с их визуальной сложностью и специфическими особенностями.

Методологически исследование базируется на сравнительно-аналитическом подходе, систематизирующем российские и международные разработки. Вычислительные методы классифицированы по алгоритмическим принципам: графовые (*MST*), ансамблевые, морфологические и лексико-семантические. Применена многомерная система оценки, включающая метрики текстового распознавания (*CER*, *IoU*) и пространственного разрешения неоднозначности (*Accuracy@161km*, *AUC*).

Проведен анализ вычислительной сложности алгоритмов (от $O(n)$ до $O(n^2)$) с учетом их масштабируемости для обработки картографических коллекций. Исследованы проблемные случаи: многословные топонимы и омонимичные географические названия. На основе комплексного анализа предложена интегрированная архитектура распознавания топонимов, сочетающая методы компьютерного зрения, обработки естественного языка и пространственного анализа.

Результаты и их обсуждение

Эволюция методов автоматического распознавания топонимов прошла путь от традиционных подходов ручного проектирования признаков, таких как *Stroke Width Transform* [Epshtein et al., 2010] и *Maximally Stable Extremal Regions* [Neumann, Matas, 2010; Neumann, Matas, 2012], к современным архитектурам глубокого обучения.

Уже с конца 1980-х годов исследователи начали разрабатывать алгоритмы для извлечения текста из изображений. Пионерская работа Fletcher и Kasturi [1988] представила метод, основанный на анализе связных компонентов и преобразовании Хафа, который, несмотря на универсальность, не мог эффективно работать с пересечениями текста и графических элементов. В 1990-х годах Pierrot-Deseilligny с соавторами [1995] усовершенствовали подход, применив оптимизацию графа для реконструкции строк символов, однако их метод требовал предварительного знания используемых шрифтов. Параллельно Woodruff и Plaunt [1994] разработали систему *GIPSY*, которая не только распознавала топонимы, но и интерпретировала пространственные отношения между ними.

Новая эра в распознавании топонимов началась с работы Smith и Crane [2001], которая охватила громадный временной диапазон от античности до XIX века и включала анализ более миллиона географических ссылок. Их двухэтапная методология объединяла эвристические методы с многоуровневым анализом контекста. Gelbukh и Levachkine [2002, 2003] сфокусировались на специфических проблемах распознавания текста на картографических материалах, выявив четыре ключевых вызова и предложив комплексное использование различных источников данных и инструментов. Отечественные исследования, представленные работой Крейдлина [2006], продемонстрировали высокую эффективность контекстного анализа для



распознавания топонимов в русскоязычных текстах, несмотря на методологические ограничения. Pouderoux с соавторами [2007] представили четырехэтапный подход к распознаванию топонимов на цветных топографических картах, объединяющий сегментацию, анализ компонентов и строк, OCR-распознавание и постобработку.

Позже Zhou с соавторами [2017] представили *EAST* – эффективный и точный детектор текста на сценах, продемонстрировавший высокую производительность на различных наборах данных. Baek с соавторами [2019] разработали *CRAFT* (*Character Region Awareness for Text Detection*), способный обнаруживать символы без необходимости индивидуальной аннотации глифов. Параллельно развивались методы связывания сущностей, такие как *GENRE* [De Cao et al., 2021] и *BLINK* [Wu et al., 2020], которые показали превосходные результаты в задачах разрешения топонимов [Sevgili et al., 2022].

Однако, несмотря на значительный прогресс, существующие подходы сталкиваются с рядом фундаментальных ограничений. Первое – это проблема многословных топонимов, где отдельные слова должны быть связаны в единую семантическую единицу. Как отмечают Olson с соавторами [2024], существуют три основные проблемы: плотность текстовых меток на картах, непоследовательность размещения многословных географических названий и наличие не связанных текстовых меток между словами многословной фразы. Второе ограничение – недостаточная репрезентация мелкомасштабных и низкочастотных топонимов в существующих базах знаний. Например, *Wikipedia* содержит только около 1,5 миллиона географических объектов, в то время как *OpenStreetMap* и *GeoNames* включают более 23 и 12 миллионов объектов соответственно [Hu et al., 2022b].

Перспективным направлением преодоления этих ограничений является интеграция различных подходов в ансамблевые модели. Hu с соавторами [2021] продемонстрировали, что ансамбль голосования, объединяющий несколько индивидуальных подходов, достигает наилучших результатов на разнородных наборах данных, значительно улучшая производительность распознавания мелкомасштабных объектов. Другое направление – создание специализированных моделей для конкретных типов карт, как предложено Lenc с соавторами [Lenc et al., 2022], которые разработали модель классификации топонимов, адаптированную для исторических кадастровых карт.

Практическое применение технологий автоматического распознавания топонимов простирается далеко за пределы академических исследований. В сфере пространственной гуманитаристики эти технологии открывают новые возможности для анализа исторической мобильности, культурных трансформаций и языковых изменений [Gregory et al., 2015]. В управлении бедствиями автоматическое распознавание топонимов в сообщениях социальных медиа помогает оперативно локализовать пострадавшие районы [Zhang et al., 2021]. В эпидемиологии географическая привязка позволяет отслеживать пространственные паттерны распространения заболеваний и планировать профилактические меры [Scott et al., 2019].

Анализируя современное состояние исследований в области автоматического распознавания топонимов, следует отметить определенный дисбаланс между отечественными и зарубежными научными работами. Хотя российские исследователи вносят ценный вклад в развитие данного направления, количество и масштаб отечественных публикаций заметно уступают зарубежным аналогам.

А.В. Дмитриев [2025] систематически привел методологические подходы, предложенные отечественными и зарубежными учеными, к автоматическому распознаванию топонимов в период с конца 1980-х до середины 2010-х годов. Автор сопоставляет их эффективность, выявляет ограничения и проблемы, которые требуют дальнейшего исследования.

В работах Д.С. Баталова и И.В. Бессмертного [2023], а также О. Аналаковой и О. Еллыевой [2024] рассматриваются возможности применения искусственного интеллекта в изучении топонимики и выявлении закономерностей на примере нескольких

статей и проектов, в которых были использованы методы анализа геопространственных данных, классификации и кластеризации геоданных, машинного и глубокого обучения для распознавания и сопоставления топонимов.

А.В. Вицентий, В.В. Диковицкий и М.Г. Шишаев [2020] разработали информационную технологию, в которой используются методы синтаксического, морфологического и семантического анализа для извлечения топонимов из текстов, а также современные веб-геосервисы для их геокодирования и визуализации. На первом этапе собирается корпус документов для анализа, на втором анализируются тексты на основе семантических и синтактико-морфологических методов и обрабатываются результаты. На последнем этапе происходит геокодирование выявленных объектов и синтез цифровой карты. Точность автоматического анализа составляет 66 %.

В работе К.К. Боярского, Е.А. Каневского и Д.А. Буториной [2019] для выявления топонимов был использован парсер *SemSin*. В результате работы данного парсера в xml-файле хранится информация о семантических, грамматических и синтаксических характеристиках слов в предложении. Авторы отмечают, что при извлечении топонимов в русскоязычных текстах часто приходится сталкиваться с проблемой семантической омонимии. В качестве решения предлагается учитывать маркерные слова, которые помогают определить класс слова с точностью не менее 95 %.

При этом необходимо отметить, что в то время как на международном уровне наблюдается интенсивное развитие методов распознавания топонимов, подкрепленное масштабными исследовательскими проектами и обширными наборами данных, российские разработки, несмотря на их методологическую ценность, представлены в меньшем объеме. Эта диспропорция создает потребность в интенсификации отечественных исследований и более активной интеграции российского научного сообщества в международное пространство разработки автоматизированных систем распознавания топонимов.

Сравнительно-сопоставительный анализ отечественных и зарубежных работ позволил выявить и систематизировать ключевые задачи, **методологические этапы** и используемый инструментарий при автоматическом распознавании топонимов на картах и в текстах (табл. 1).

Таблица 1
Table 1

Сравнение методологических подходов к обработке топонимов
Comparison of methodological approaches to toponym processing

Авторы	Задача	Методологические этапы	Особенности подхода	Используемые программные и математические решения
Lenc et al. [2022]	Извлечение топонимов с исторических кадастровых карт	1. Обнаружение топонимов 2. Классификация топонимов 3. Распознавание текста (OCR)	Разделение топонимов на муниципальные (печатные) и общие (рукописные)	Faster R-CNN, EAST, HP-FCN, YOLOv5, Tesseract OCR
Milleville et al. [2020]	Улучшение точности распознавания топонимов на исторических топографических картах	1. Предварительная обработка и геопривязка 2. Сжатие и разделение на фрагменты 3. Распознавание текста 4. Сопоставление с базами топонимов	Изучение влияния качества сжатия и размера фрагментов на точность распознавания	Azure Read API, методы компьютерного зрения, газеттиры



Окончание таблицы 1
End of the table 1

Olson et al. [2024]	Поиск многословных топонимов на исторических картах	1. Получение текстовых меток из карты 2. Построение минимального остоного дерева 3. Поиск многословного топонима в дереве	Использование MST для определения отношений между одиночными словами	Алгоритм Прима для MST, функция стоимости ребра на основе визуальных атрибутов
Hu et al. [2023]	Разрешение неоднозначности топонимов	1. Группировка координатных оценок с помощью DBSCAN 2. Выбор крупнейшего кластера 3. Использование центроида выбранного кластера как результат	Ансамблевый подход на основе «голосования» нескольких методов	DBSCAN, ансамбль из 7 алгоритмов (GENRE, BLINK, LUKE, CamCoder, Edinburgh Geoparser, SHS, CBH)
Schlegel [2021]	Автоматическое извлечение меток с крупномасштабных исторических карт	1. Обнаружение текста 2. Распознавание текста 3. Измерение сходства строк 4. Приблизительная геопривязка	Комбинирование инструментов обнаружения, распознавания и измерения сходства для повышения точности	Strabo, TensorFlow, OpenCV, KAZE, расстояние Левенштейна
Gritta et al. [2018]	Геокодирование с использованием векторов карты	1. Извлечение характеристик из текста 2. Обучение модели CNN 3. Классификация местоположения	Использование как локальных, так и глобальных признаков из контекста	CNN, предварительно обученная на 1,4 млн геопривязанных статей Википедии
Sekido [2021]	Создание исторических ГИС-материалов для исследований	1. Реорганизация данных в соответствии с целью 2. Преобразование топонимов в географическую информацию 3. Выбор базовой карты 4. Создание карт	Интеграция исторических данных в ГИС	India Place Finder, ГИС-инструменты, исторические карты
Weissenbacher et al. [2015]	Распознавание топонимов в научных статьях	1. Выбор кандидата топонима по стране 2. Выбор по населению при неразрешимой неоднозначности	Использование эвристик, основанных на особенностях научных публикаций	Heuristic-based rules
Вицентий и Шишаев [2021]	Извлечение геоатрибутированных сущностей из текстов на естественном языке	1. Нейросетевой анализ 2. Синтаксико-морфологический анализ 3. Лексико-семантический анализ 4. Геокодирование и геовизуализация	Гибридная технология с учетом «прямого» и «описательного» представления геоатрибутированных сущностей	Библиотека Natasha (на основе BERT), Yargy-парсер с алгоритмом Earley, лексико-семантические шаблоны на LSPL, GIS

Табл. 1 дает возможность оценить релевантность и специфические особенности данных подходов в аспекте геоинформационного, географического, топонимического и математического анализа.

С точки зрения геоинформатики наиболее полными представляются подходы, учитывающие пространственный контекст. Метод Ну с соавторами [2023] с использованием *DBSCAN* для кластеризации географических координат эффективно работает с пространственной неопределенностью. Преимущество – способность учитывать множественные гипотезы о местоположении, недостаток – зависимость от параметров кластеризации. Методология Olson с соавторами [2024] игнорирует геопространственные отношения между различными топонимами на карте, что ограничивает ее применимость для сложных случаев пространственного контекста. Подход Sekido [2019] отличается фокусом на историческую преемственность координатных систем, что критично для интеграции старых и современных данных. Значительный вклад в геоинформационную составляющую вносит методология Вицентия и Шишаева [2021], предлагающая четкую последовательность геокодирования и визуализации структурированных данных, с возможностью синтезировать картосхемы почти в режиме реального времени, что особенно актуально для систем поддержки принятия решений.

Географический аспект наиболее полно учтен в методологии Milleville с соавторами [2020], где геопривязка является начальным этапом процесса, что позволяет ограничить область поиска кандидатов для сопоставления топонимов. Подход Gritta с соавторами [2018] выигрывает благодаря учету глобальных географических контекстов и популяционных данных при разрешении неоднозначностей. Существенный недостаток методов, представленных Lenc с соавторами [2022] и Schlegel [2021], – игнорирование географической иерархии объектов (страна – регион – город), что может приводить к ошибкам при омонимии. Методы с большим весом текстовых признаков и малым весом пространственных особенностей (например, в работе Weissenbacher et al., 2015) малоэффективны для топографических и исторических карт с плотным расположением объектов. Работа Вицентия и Шишаева [2021] выделяется введением понятия геоатрибутированной сущности (ГаС), которое расширяет традиционное понимание топонима, включая «описательные» пространственные объекты, что особенно важно для анализа транспортно-логистических систем и региональных пространственных структур.

С позиции топонимики наиболее обоснованными представляются подходы, учитывающие лингвистические особенности топонимов. Подход Schlegel [2021] с использованием расстояния Левенштейна учитывает вариативность написания, но не отражает топонимическую типологию и таксономические отношения. Методология Olson et al. [2024] для многословных топонимов учитывает структурные особенности (регистр, интервалы, размер шрифта), что соответствует топонимической практике. Вместе с тем ни один из методов не учитывает этимологические и культурно-исторические аспекты топонимов, что критично для исторических карт с изменившимися названиями. Работа Lenc с соавторами [2022] с разделением на муниципальные и общие топонимы представляет удачный пример таксономического подхода, но требует дальнейшей детализации. Методология Вицентия и Шишаева [2021] отличается использованием лексико-семантических шаблонов, которые учитывают грамматические особенности и валентности лексем в русском языке, что позволяет точнее классифицировать топонимические объекты и определять их семантические свойства, особенно для сложных лексических конструкций и описательных топонимов.

Математически наиболее обоснованными являются подходы, использующие формальные модели для оценки сходства и кластеризации. Метод Ну с соавторами [2023] с применением *DBSCAN* демонстрирует робастность к выбросам, но чувствителен к выбору



параметров. Алгоритм *MST* в работе Olson с соавторами [2024] обеспечивает оптимальное связывание компонентов с вычислительной сложностью $O(n^2)$, что недостаточно эффективно для карт с большим количеством объектов. CNN-подход Gritta с соавторами [2018] обеспечивает гибкость модели и способность к обобщению, но требует значительных вычислительных ресурсов и обучающих данных. Наиболее эффективное соотношение сложности и точности демонстрирует ансамблевый подход Ну с соавторами [2023], совмещающий преимущества разных моделей и снижающий их индивидуальные недостатки. Работа Вицентия и Шишаева [2021] представляет математически обоснованный гибридный подход, сочетающий преимущества нейросетевых, правило-ориентированных и шаблонных методов, с эмпирически определенными параметрами (например, размер «окрестности» $[-1; +1]$ для контекстного анализа), что обеспечивает достаточно высокую точность (66 %) при обработке реальных текстовых корпусов.

Таким образом, наиболее комплексными и эффективными представляются гибридные подходы, сочетающие пространственный контекст, лингвистические особенности топонимов и робастные алгоритмы кластеризации. Главные ограничения существующих методов – недостаточный учет исторических изменений топонимов, отсутствие интеграции культурно-географического контекста и высокие требования к вычислительным ресурсам для методов глубокого обучения.

Следующим этапом в нашем анализе стало более глубокое, детальное сравнительное исследование вычислительного аппарата методов распознавания и обработки топонимов в российских и зарубежных трудах.

Представленные ниже таблицы демонстрируют комплексную картину методологических и алгоритмических подходов к проблемам распознавания и геопространственной интерпретации топонимов. При детальном рассмотрении этого вычислительного ландшафта выявляется ряд критических закономерностей и тенденций.

В области метрик оценки качества распознавания текста (табл. 2) наблюдается четкая дихотомия между символьно-ориентированными и геометрическими подходами. Символьные метрики, такие как *CER*, демонстрируют значительную вариативность результатов (от 0,061 до 0,512 у Lenc et al., 2022), что свидетельствует о существенном влиянии типа текста на качество распознавания. Примечательно, что коммерческие системы показывают более стабильные результаты (0,217–0,301) по сравнению с открытыми (0,284–0,351). Это указывает на потенциальные преимущества проприетарных алгоритмов в обработке гетерогенных текстовых данных на картографических материалах.

Геометрические метрики, в частности *IoU*, обнаруживают меньшую вариативность (0,516–0,630), что может объясняться более универсальным характером пространственного соответствия по сравнению с лингвистическими особенностями текста. Многоуровневый подход к оценке *IoU* (на уровнях 0,5 и 0,75), используемый Lenc с соавторами [2022], представляется методологически более обоснованным, поскольку позволяет дифференцировать «приемлемое» и «точное» соответствие, что критично для картографических приложений.

Анализируя метрики разрешения неоднозначности топонимов (табл. 3), следует отметить концептуальное различие между дистанционно-ориентированными и вероятностными подходами. Метрика *Accuracy@161km*, предложенная Ну с соавторами [2023], устанавливает четкий пространственный порог приемлемости, что соответствует практическим требованиям в геопространственных приложениях. При этом интегральная метрика *AUC* с логарифмическим масштабированием позволяет более тонко оценивать распределение ошибок, придавая меньший вес экстремальным выбросам. Это особенно важно при работе с историческими картами, где точная геопривязка зачастую невозможна.

Таблица 2
Table 2

Метрики оценки качества распознавания текста
Metrics for evaluating text recognition quality

Тип метрики	Исследование	Название метрики	Формальное определение	Диапазон значений в исследовании
Символьный уровень	Milleville et al. [2020]	Показатель ошибок по символам (CER)	Расстояние Левенштейна между предсказанным текстом и меткой, деленное на длину метки	0,217–0,301 (коммерческая система) 0,284–0,351 (открытая система)
	Lenc et al. [2022]	Показатель ошибок по символам (CER)	То же	0,061–0,512 (в зависимости от типа текста)
Геометрическое соответствие	Milleville et al. [2020]	Пересечение над объединением (IoU)	Площадь пересечения / площадь объединения	0,516–0,630
	Lenc et al. [2022]	IoU@50, IoU@75	Доля предсказаний с IoU более 0,5 и 0,75	IoU@50: 0,654–0,872 IoU@75: 0,539–0,806
Совокупные метрики	Lenc et al. [2022]	Average Precision (AP)	Площадь под кривой precision/recall	0,271–0,778
		F1-мера	$2 * (\text{precision} * \text{recall}) / (\text{precision} + \text{recall})$	0,162–0,871
		Avg AP	Среднее AP при IoU от 0,5 до 0,95 с шагом 0,05	2,78–46,7

Таблица 3
Table 3

Метрики оценки разрешения неоднозначности топонимов
Metrics for evaluating toponym disambiguation

Тип метрики	Исследование	Название метрики	Формальное определение	Особенности применения
Расстояние	Hu et al. [2023]	Accuracy@16 km	Процент ошибок геокодирования, которые меньше 100 миль (161 км)	Оценка доли правильно геокодированных топонимов с допустимой погрешностью
	Hu et al. [2023]	Средняя ошибка (ME)	Среднее расстояние между предсказанными и истинными координатами	Количественная оценка среднего отклонения в километрах
Интегральные	Hu et al. [2023]	Площадь под кривой (AUC)	$\int_{i=1 \text{ to } N} [\ln(x_i + 1) / \ln(20039)] dx$	Учитывает распределение ошибок по всем топонимам с логарифмическим масштабированием
Точность	Вицентий и Шишаев [2021]	Средняя точность автоматического анализа	Доля правильно распознанных геоатрибутированных сущностей	66 % на тестовом корпусе из 24322 слов



В исследовании Hu с соавторами [2023] метрика *Accuracy@161km* была применена для оценки эффективности их ансамблевого метода при разрешении крайне неоднозначных топонимов. Например, для топонима "Victoria Park", имеющего 589 возможных вариантов в базе *GeoNames*, их метод смог правильно определить, что речь идет о парке в Лондоне, основываясь на контексте: "The Clash – White Riot (Live 1978 Victoria Park, London): via @YouTube Let's start our shift!". Аналогично для аббревиатуры "SA" с 58 вариантами в *GeoNames* система правильно выбрала Южную Африку. Несмотря на высокую вычислительную сложность $O(n \log n)$, метод успешно обработал 98300 топонимов из 12 различных наборов данных.

Примечательно, что в работе Вицентия и Шишаева [2021] используется более обобщенная метрика средней точности (66 %), что отражает их фокус не только на топонимах, но и на более широком понятии геоатрибутированных сущностей, включающем пространственные отношения между объектами.

Табл. 4 представляет алгоритмы обработки и классификации топонимических данных и отражает многоуровневую структуру современных подходов к распознаванию и интерпретации географических названий. Примечательна методологическая дифференциация между алгоритмами, оперирующими в текстовом, визуальном и геопространственном доменах.

Таблица 4
Table 4

Алгоритмы обработки и классификации топонимических данных
Algorithms for processing and classification of toponymic data

Тип алгоритма	Исследование	Название алгоритма	Формализация/ Параметры	Вычислительная сложность	Применение
Графовые методы	Olson et al. [2024]	Минимальные остовные деревья (MST)	$\text{cost}(l_i, l_j) = d(l_i, l_j) * (1 + h(l_i, l_j)) * (1 + a(l_i, l_j)) * (1 + c(l_i, l_j))$	$O(n^2)$ для построения MST $O(n)$ для поиска	Связывание отдельных слов в многословные топонимы
Строковые алгоритмы	Milleville et al. [2020]	Сопоставление подстрок	Объединение по наибольшей общей подстроке	$O(m*n)$, где m и n – длины строк	Объединение фрагментов текста (например, "Sint-Pie" и "ikt-Pieter")
Морфологические операции	Lenc et al. [2022]	Анализ связанных компонентов	Медианный фильтр + бинаризация + морфологическая дилатация	$O(n)$, где n – количество пикселей	Предварительная обработка изображений карт
Лексические шаблоны	Вицентий и Шишаев [2021]	Лексико-семантические шаблоны	$LSP1 = \{A \text{ or } P\}$ $N(\{aw\}) < A = N,$ $P = N, A = P >$	Зависит от сложности шаблона	Распознавание сложных топонимических конструкций
Плотностная кластеризация	Hu et al. [2023]	DBSCAN	$\text{minPts} = 2$ $\text{eps} = 10 \text{ км}$	$O(n \log n)$	Группировка координатных оценок разных методов
Дескрипторы изображений	Lenc et al. [2022]	KAZE-дескрипторы	Евклидово расстояние для сопоставления	$O(n \log n)$	Классификация топонимов на печатные и рукописные
Нейросетевые	Вицентий и Шишаев [2021]	BERT (библиотека Natasha)	Предобучение на русскоязычных текстах	Зависит от размера модели	Распознавание 97–98 % именованных сущностей
Комбинированные	Hu et al. [2023]	Ансамблевый подход	7 алгоритмов с дифференцированными весами	Сумма сложностей всех методов	Повышение робастности распознавания

Olson с соавторами [2024] продемонстрировали эффективность этого подхода на примере распознавания сложных многословных топонимов, таких как "Sault Ste. Marie". Алгоритм сначала находит слово "Sault", а затем для каждой текстовой метки, соответствующей этому слову, ищет связанные с ней в графе метки со словами "Ste." и "Marie". Также система успешно обрабатывает случаи множественной омонимии для топонима "Santa Maria", различая между "(Santa Maria, California)", "(Santa Maria, Bulacan)" и другими вариантами. Результаты тестирования показали, что их метод правильно связывает 82,61 % многословных топонимов. Однако квадратичная временная сложность $O(n^2)$ делает этот подход проблематичным для масштабных картографических коллекций с высокой плотностью текстовых аннотаций, что подтверждается ограниченным тестированием на наборе из 31500 топонимов.

В группе алгоритмов текстовой обработки наиболее методологически обоснованным представляется подход Olson с соавторами [2024] с использованием минимальных остовных деревьев. Функция стоимости ребер MST элегантно интегрирует четыре критических параметра связности текстовых меток: пространственное расстояние, отношение высот, угловое расхождение и капитализацию. Такая многомерная формализация преодолевает традиционные ограничения простых эвристик, основанных только на дистанции. Однако квадратичная временная сложность $O(n^2)$ делает этот подход проблематичным для масштабных картографических коллекций с высокой плотностью текстовых аннотаций.

Контрастирующим по вычислительной эффективности выступает метод Lenc с соавторами [2022], основанный на морфологических операциях с линейной сложностью $O(n)$. Последовательность «медианный фильтр → бинаризация → морфологическая дилатация → анализ связных компонентов» представляет собой классический конвейер обработки изображений, адаптированный для картографической специфики. Такой подход оптимален для предварительной обработки, но его точность сильно зависит от качества исходных изображений и четкости разделения текстовых и нетекстовых элементов карты (рис. 1).

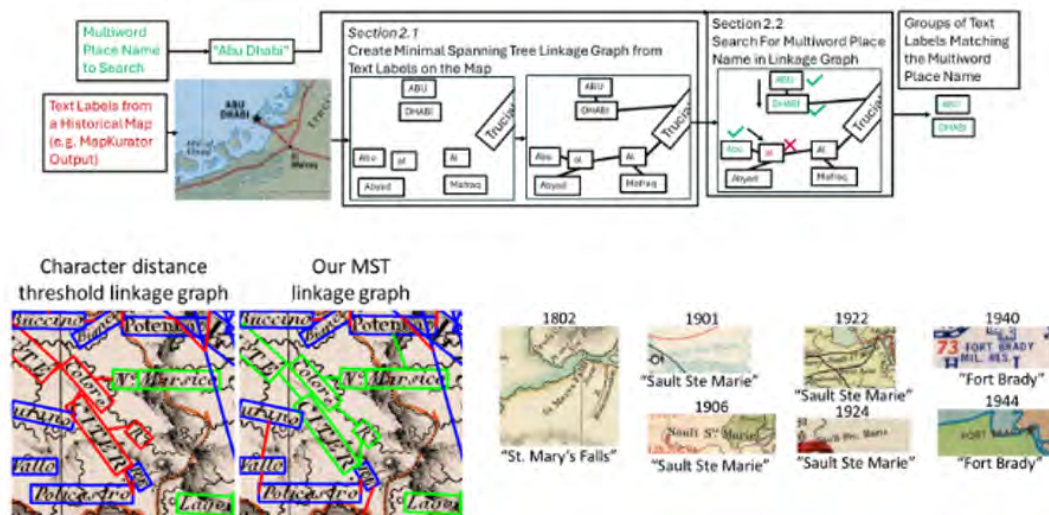


Рис. 1. Сравнение методов связывания многословных топонимов: пороговое расстояние между символами и алгоритм минимального остовного дерева (MST) на примере исторических карт с топонимом "Sault Ste. Marie" разных периодов [Olson et al., 2024]

Fig. 1. Comparison of methods for linking multi-word toponyms: character distance threshold and minimum spanning tree (MST) algorithm using historical maps with the toponym "Sault Ste. Marie" from different periods [Olson et al., 2024]

Строковые алгоритмы, представленные в работе Milleville с соавторами [2020], занимают промежуточное положение по вычислительной сложности $O(m*n)$ и демонстрируют прагматичный подход к проблеме фрагментированных текстовых меток. Milleville с соавторами [2020] успешно применили этот метод для объединения фрагментов "Sint-Pie" и "ikt-Pieter" в единый топоним "Sint-Pieter" на основе их общей подстроки (рис. 2).

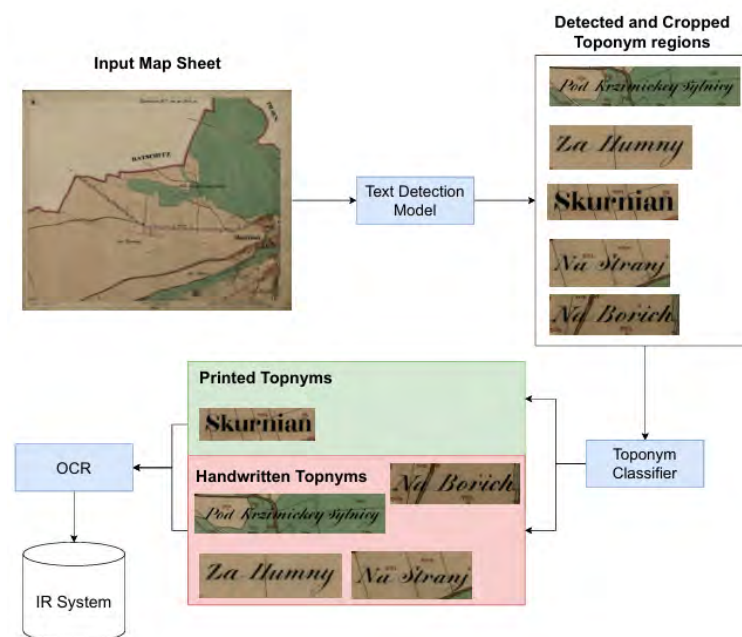


Рис. 2. Общий пайплайн по распознаванию топонимов на исторических картах, представленный в работе Lenc и соавторов и реализующий алгоритм анализа связанных компонент [2022]

Fig. 2. General pipeline for toponym recognition on historical maps, presented in the work of Lenc et al. [2022], implementing the connected component analysis algorithm

Однако авторы также отмечают проблемные случаи, такие как слова "Oude Leieput" и "Hbg Karper", где возникли сложности с определением правильного способа объединения соседних слов. Чувствительность метода к качеству исходных изображений подтверждается экспериментами на фрагментах размером от 1000×1000 до 2500×2500 пикселей с различным уровнем сжатия.

Особый интерес представляют лексико-семантические шаблоны Вицентия и Шишова [2021], формализующие структурные и грамматические отношения в топонимах. Шаблон $LSP1 = \{A \text{ or } P\} N(\{aw\}) <A = N, P = N, A = P>$ кодифицирует морфосинтаксические паттерны, специфичные для русского языка. Исследователи успешно применили свой подход как к прямым топонимам («Хибины», «Мурманск»), так и к сложным описательным геоатрибутированным сущностям, таким как «железнодорожная линия: станция Выходной – мостовой переход через р. Тулома – станция Мурманши 2 – станция Лавна». Используя лексико-семантические шаблоны вида $LSP1 = \{A \text{ or } P\} N(\{aw\}) <A = N, P = N, A = P>$, авторы не только распознают сами топонимы, но и их атрибуты. Например, из фразы «Новая асфальтовая автодорога будет проложена между Мурманском и Белокаменкой» выделяется геоатрибутированная сущность «автодорога» класса «автомобильная дорога», а также значения атрибутов «тип покрытия» и «прочие характеристики». Тестирование на корпусе из 24322 слов показало среднюю точность автоматического анализа 66 %. Этот подход потенциально наиболее адаптивен к морфологически богатым языкам, где простые строковые алгоритмы неэффективны. Однако вычислительная сложность этого метода сильно зависит от количества и детализации шаблонов, что затрудняет его масштабирование.

Переходя к алгоритмам классификации, необходимо отметить методологическую зрелость DBSCAN [Hu et al., 2023] в контексте пространственной кластеризации. Параметризация $minPts = 2$, $eps = 10$ км оптимизирована для географических данных и обеспечивает баланс между чувствительностью к локальным скоплениям и устойчивостью к выбросам. Логарифмическая сложность $O(n \log n)$ делает этот алгоритм масштабируемым для больших наборов данных (рис. 3).

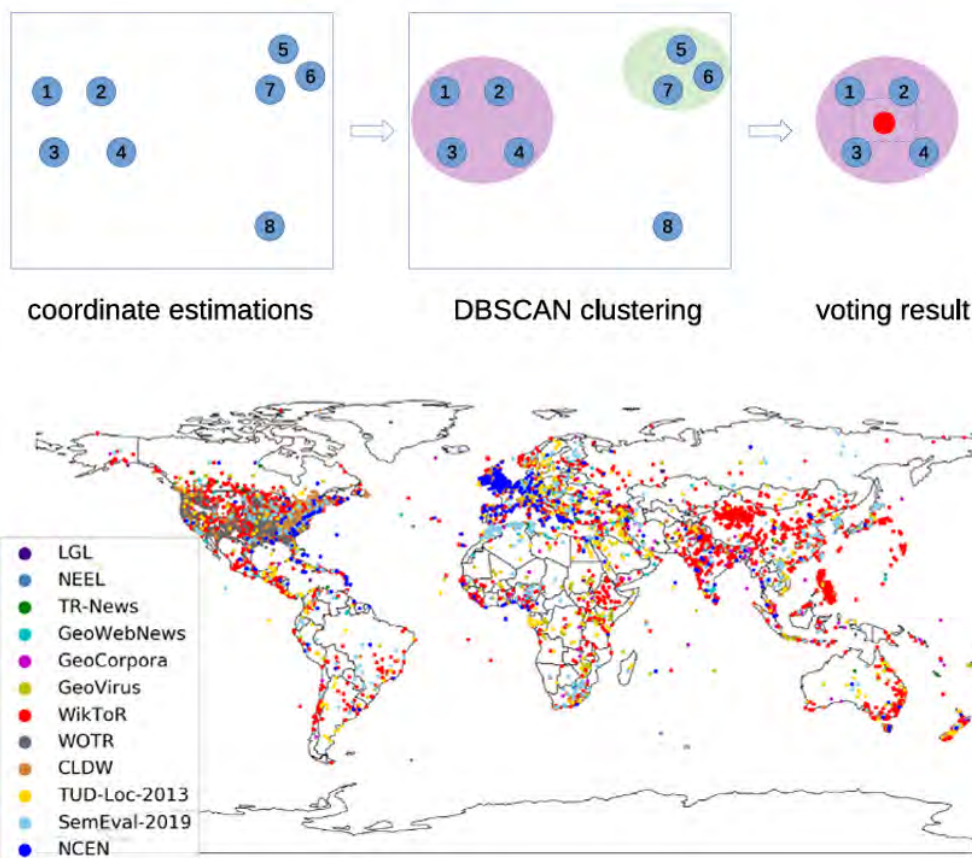


Рис. 3. Рабочий процесс голосующего подхода (выше) и результат применения ансамблевого алгоритма – пространственное распределение 98300 топонимов в 12 наборах данных [Hu et al., 2023]
Fig. 3. Workflow of the voting approach and its results – spatial distribution of the 98300 toponyms in the 12 datasets [Hu et al., 2023]

Использование *KAZE*-дескрипторов (Lenc et al., 2022) для классификации топонимов демонстрирует важный концептуальный сдвиг от текстового к визуальному домену. Применение инвариантных к масштабу и повороту дескрипторов позволяет эффективно дифференцировать печатные и рукописные топонимы, что критично для исторических карт с гетерогенными стилями надписей. Однако точность этого подхода может снижаться при низком контрасте или значительных перекрытиях текста с другими элементами карты.

Нейросетевые методы, представленные библиотекой *Natasha* на базе *BERT* [Виценций, Шишаев, 2021], отражают современную тенденцию к использованию предобученных языковых моделей. Высокая точность распознавания именованных сущностей (97–98 %) обеспечивается не только архитектурой трансформера, но и специализированным предобучением на русскоязычных текстах. Это подчеркивает важность лингвистической специфики при обработке топонимов.

Наиболее методологически продвинутым представляется ансамблевый подход Ну с соавторами [2023], интегрирующий 7 разнородных алгоритмов с дифференцированными весами. Такая архитектура обеспечивает максимальную робастность за счет компенсации слабостей отдельных методов, что особенно ценно при обработке исторических карт с вариативным качеством. Однако суммарная вычислительная сложность всех методов и необходимость их координации делают этот подход ресурсоемким.

Сравнивая все алгоритмы в континууме «вычислительная эффективность ↔ точность распознавания», можно отметить общую тенденцию к усложнению методологического аппарата при повышении требований к качеству распознавания. Графовые методы (*MST*) и комбинированные подходы (ансамбли) представляют верхний предел точности



при высокой вычислительной стоимости, тогда как морфологические операции и строковые алгоритмы обеспечивают лучшую масштабируемость при потенциально большем количестве ошибок.

Интегрируя данные наблюдения, можно заключить, что оптимальной стратегией обработки топонимических данных является многоуровневый каскадный подход с динамическим выбором алгоритмов в зависимости от типа карты, плотности аннотаций и требуемой точности. Прогрессивное усложнение анализа – от простых морфологических операций и строковых алгоритмов к графовым и ансамблевым методам – позволяет эффективно балансировать между вычислительными ресурсами и качеством распознавания.

Геопространственные алгоритмы (табл. 5) демонстрируют методологическую зрелость в области геопривязки и разрешения неоднозначности.

Таблица 5
Table 5

Геопространственные алгоритмы
Geospatial algorithms

Функциональность	Исследование	Метод	Описание	Особенности
Геопривязка	Milleville et al. [2020]	Каскадный процесс	Обнаружение границ + OCR + преобразование координат	Усреднение для снижения ошибок
Пространственные запросы	Milleville et al. [2020]	Регион интереса	Квадратная рамка круговой области с радиусом 2 км	Ограничение области поиска
Контекстный анализ	Вицентий и Шишаев [2021]	«Скольльзящее окно»	Размер окна [-1; +1]	Учет контекста левого и правого предложений
Разрешение неоднозначности	Hu et al. [2023]	Выбор центроида крупнейшего кластера	Центроид координатных оценок в выбранном кластере	Робастность к выбросам

Каскадный процесс Milleville с соавторами [2020] с усреднением для снижения ошибок представляет собой прагматичный подход к проблеме неточности исторических карт. Метод «скользящего окна» Вицентия и Шишаева [2021] значительно расширяет контекстное понимание топонимов, хотя ограничивается текстовым измерением без учета пространственных отношений между объектами.

Центроидный метод Hu с соавторами [2023] для разрешения неоднозначности интегрирует множественные гипотезы о местоположении и обеспечивает робастность к выбросам. В их работе продемонстрировано, как ни один из отдельных алгоритмов в ансамбле не мог правильно разрешить все тестовые случаи с неоднозначными топонимами ("SA", "False River", "Victoria Park", "Mount Sheridan"), но их комбинация с последующей кластеризацией DBSCAN ($minPts = 2$, $eps = 10$ км) успешно справилась с задачей. Этот подход обладает временной сложностью $O(n \log n)$, что делает его применимым к значительным объемам данных.

Анализ вычислительной эффективности (табл. 6) выявляет значительные различия в масштабируемости рассмотренных подходов. Временная сложность варьируется от $O(n^2)$ для методов на основе MST до линейной $O(n)$ для анализа связанных компонентов. Это напрямую влияет на практическую применимость методов к крупномасштабным картографическим коллекциям. Примечательно, что подходы с более низкой вычислительной сложностью [Вицентий, Шишаев, 2021; Lenc et al., 2022] обрабатывают и меньшие объемы данных, что может указывать на компромисс между вычислительной эффективностью и масштабируемостью.

Наиболее вычислительно интенсивный подход Olson с соавторами [2024] с квадратичной сложностью $O(n^2)$ был применен к набору из 31500 топонимов (*WikToR*), но авторы отмечают его неэффективность для карт с большим количеством объектов. В противоположность этому, подход Lenc с соавторами [2022] с линейной сложностью $O(n)$ позволил обработать 800 картографических листов с 3700 регионами, но его эффективность ограничивается зависимостью от качества исходного изображения.

Таблица 6
Table 6

Сравнение вычислительной эффективности
Comparison of computational efficiency

Исследование	Временная сложность	Пространственная сложность	Обработанный объем данных	Ограничения
Olson et al. [2024]	$O(n^2)$ для MST $O(n)$ для поиска	$O(n)$ для хранения ребер	31500 топонимов (<i>WikToR</i>)	Неэффективно для карт с большим количеством объектов
Hu et al. [2023]	$O(n \log n)$ для DBSCAN	Память для результатов всех методов	98300 топонимов из 12 наборов	Высокие требования к вычислительным ресурсам
Milleville et al. [2020]	Зависит от размера фрагмента	Хранение промежуточных результатов	Фрагменты 1000×1000 – 2500×2500	Чувствительность к качеству сжатия
Lenc et al. [2022]	$O(n)$ для анализа связанных компонентов	Зависит от размера изображения	800 картографических листов	Зависимость от качества исходного изображения
Вицентий и Шишаев [2021]	Зависит от размера корпуса	Хранение контекстных окрестностей	Корпус из 24322 слов	Зависимость от качества разработанных правил и словарей

Наиболее масштабный анализ провели Hu с соавторами [2023], обработав 98300 топонимов из 12 различных наборов данных с использованием алгоритма DBSCAN сложности $O(n \log n)$, но это потребовало значительных вычислительных ресурсов и хранения промежуточных результатов всех методов в ансамбле.

Интегрируя наблюдения по всем аспектам анализа вычислительной эффективности автоматического распознавания топонимов, можно выделить три ключевые методологические тенденции в современных подходах к обработке топонимических данных (табл. 7):

1. Комплексность метрик оценки – переход от одномерных метрик к многоаспектным системам оценки, учитывающим как текстовые, так и пространственные характеристики топонимов. Это наглядно демонстрируют работы Lenc et al. с многоуровневым подходом к оценке *IoU* и Hu с соавторами с использованием метрики *Accuracy@161km* вместе с площадью под кривой (*AUC*).

2. Гибридизация алгоритмических подходов – интеграция классических методов обработки текста с геопространственными алгоритмами и методами машинного обучения. Яркий пример – ансамблевый подход Hu с соавторами [2023], который успешно применил комбинацию различных алгоритмов для разрешения сложных случаев неоднозначности топонимов, таких как "*Victoria Park*" с 589 возможными вариантами.

3. Контекстуализация анализа – растущее внимание к контекстному окружению топонимов. Метод «скользящего окна» Вицентия и Шишаева [2021] с размером окна $[-1; +1]$ позволил успешно распознавать сложные описательные геоатрибутированные сущности, а пространственный подход Olson с соавторами [2024] учитывает не только расстояние между словами, но и их взаимное расположение, высоту и капитализацию.



Наиболее перспективным направлением развития представляется создание адаптивных систем, способных динамически выбирать оптимальный набор алгоритмов и метрик в зависимости от характеристик обрабатываемых данных, исторического периода карт и задач конкретного исследования.

На рис. 4 представлена общая итоговая концепция, демонстрирующая своего рода пайплайн автоматической обработки топонимов на картах и в текстах.

Таблица 7
Table 7

Типы проблем распознавания
Types of recognition problems

Исследование	Алгоритм	Вычислительная сложность	Практический пример	Результат
Многословные топонимы				
Olson et al. [2024]	Минимальные остовные деревья (MST)	$O(n^2)$	"Sault Ste. Marie" – алгоритм сначала ищет "Sault", затем связанные с ним "Ste." и "Marie"	82,61 % правильно связанных многословных топонимов
Milleville et al. [2020]	Объединение по наибольшей общей подстроке	$O(m*n)$	"Sint-Pieter" распознан путем объединения фрагментов "Sint-Pie" и "ikt-Pieter" на основе общей подстроки	Точность зависит от размера фрагмента и качества изображения
Виценти и Шишаев [2021]	Лексико-семантические шаблоны	Зависит от сложности шаблона	«железнодорожная линия: станция Выходной – мостовой переход через р. Тулома – станция Мурманши 2 – станция Лавна»	66 % средняя точность автоматического анализа
Неоднозначные топонимы				
Hu et al. [2023]	Ансамблевый подход с DBSCAN	$O(n \log n)$	"Victoria Park" – из 589 возможных вариантов в GeoNames правильно определен как парк в Лондоне	Accuracy@161km: 0,86
Hu et al. [2023]	Ансамблевый подход с DBSCAN	$O(n \log n)$	"SA" – из 58 вариантов правильно определена как Южная Африка	Правильное разрешение в контексте с критической информацией
Виценти и Шишаев [2021]	«Скользющее окно» с размером [-1; +1]	Зависит от размера корпуса	«Новая автодорога будет проложена между Мурманском и Белокаменкой»	Распознаны топонимы и их пространственные отношения
Визуальные и форматные отличия				
Lenc et al. [2022]	Морфологические операции и KAZE-дескрипторы	$O(n)$ для связанных компонентов, $O(n \log n)$ для дескрипторов	Классификация на муниципальные (печатные) и общие (рукописные) топонимы	99,2 % точность классификации
Milleville et al. [2020]	Каскадный процесс	Зависит от размера фрагмента	"Oude Leieput" и "Hbg Karper" - проблема определения, какие соседние слова должны быть объединены	Проблемы с нечетким поиском: "ent" не возвращает "Gent"



Рис. 4. Комплексная архитектура автоматического распознавания топонимов: от данных к результатам
Fig. 4. Comprehensive architecture of automatic toponym recognition: from data to results

Представленный обзор демонстрирует значительный прогресс и одновременно сохраняющиеся вызовы в области автоматического распознавания топонимов на картографических материалах. Анализ вычислительных подходов выявляет четкую тенденцию к созданию гибридных систем, интегрирующих методы компьютерного зрения, обработки естественного языка и пространственного анализа.

Сравнительная оценка эффективности различных алгоритмов показывает существенные методологические расхождения между графовыми методами, ансамблевыми подходами, морфологическими операциями и лексико-семантическими методами. При этом наблюдается закономерный компромисс между вычислительной сложностью и точностью распознавания.

Метрические системы оценки эволюционируют от одномерных показателей к многоаспектным критериям, учитывающим как текстовые, так и пространственные характеристики. Особенно перспективным представляется использование метрик, специфичных для геопространственных задач (*Accuracy@161km*), в сочетании с классическими оценками текстового распознавания (*CER*).

Принципиальной проблемой остается обработка многословных топонимов, что обусловлено не только визуальными особенностями карт, но и лингвистической спецификой географических названий. Методы на основе минимальных остовных деревьев демонстрируют многообещающие результаты, однако их квадратичная вычислительная сложность ограничивает применимость для масштабных коллекций.

Разрешение неоднозначности топонимов эффективно решается ансамблевыми методами с пространственной кластеризацией, но требует значительных вычислительных ресурсов. Примечательно, что ни один из индивидуальных алгоритмов не достигает приемлемой точности для сложных случаев многозначности.

В контексте визуальных и форматных различий топонимов морфологические операции в сочетании с дескрипторами изображений обеспечивают высокую точность классификации, однако их эффективность сильно зависит от качества исходных картографических материалов.



Заключение

Развитие автоматизированных подходов к распознаванию топонимов открывает значительные перспективы в создании интеллектуальных систем картографического наследия, способных обрабатывать обширные исторические коллекции. Проведенный анализ демонстрирует эволюцию методов от традиционных подходов к гибридным архитектурам, интегрирующим компьютерное зрение, обработку естественного языка и пространственный анализ. Будущее этого направления лежит в создании адаптивных систем, самостоятельно выбирающих оптимальные алгоритмы в зависимости от типа и качества обрабатываемых материалов, а также в разработке методов, учитывающих исторические изменения топонимов, культурный контекст и лингвистические особенности разных языков. Перспективными направлениями исследований являются также использование мультимодальных подходов для обработки карт с высоким уровнем визуальной сложности, интеграция с историческими газетными архивами и создание открытых наборов данных для тренировки специализированных моделей. Это позволит не только повысить точность распознавания сложных случаев (многословные и неоднозначные топонимы), но и создать инструменты для междисциплинарных исследований на стыке географии, истории, лингвистики и компьютерных наук.

Список литературы

- Аналакова О., Еллыева О. 2024. Использование искусственного интеллекта для анализа пространственных данных и выявления закономерностей в картографии. *Инновационная наука*, 5–2–2: 158–160.
- Баталов Д.С., Бессмертный И.В. 2023. Применение искусственного интеллекта в топонимике. Обзор методов искусственного интеллекта и научных работ. В кн.: *Географические аспекты устойчивого развития регионов. V Международная научно-практическая конференция*, Гомель, 25–26 мая 2023. Гомель, Гомельский государственный университет имени Франциска Скорины: 372–376.
- Боярский К.К., Каневский Е.А., Буторина Д.А. 2019. К вопросу о разрешении семантической омонимии топонимов в русскоязычных текстах. *Компьютерная лингвистика и вычислительные онтологии*, 3: 19–28. <https://doi.org/10.17586/2541-9781-2019-3-19-28>
- Вицентий А.В., Диковицкий В.В., Шишаев М.Г. 2020. Технология извлечения и визуализации пространственных данных, полученных при анализе текстов. *Труды Кольского научного центра РАН*, 11(8(11)): 115–119. <https://doi.org/10.37614/2307-5252.2020.8.11.012>
- Вицентий А.В., Шишаев М.Г. 2021. Технология извлечения геоатрибутированных сущностей для визуального представления пространственной связности объектов на основе автоматизированной генерации картосхем. *Труды Кольского научного центра РАН. Информационные технологии*, 12(5): 35–49.
- Дмитриев А.В. 2025. Ранний этап в методологии автоматического распознавания топонимов на географических картах и в текстах. *Региональные геосистемы*, 49(1): 128–145. <https://doi.org/10.52575/2712-7443-2025-49-1-128-145>
- Крейдлин Л.Г. 2006. Программа выделения русских индивидуализированных именных групп TagLite. В кн.: *Компьютерная лингвистика и интеллектуальные технологии. Ежегодная международная конференция «Диалог»*, Москва, 01–04 июня 2016. М., Наука: 292–297.
- Ardanuy M.C., Sporleder C. 2017. Toponym Disambiguation in Historical Documents Using Semantic and Geographic Features. *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, Göttingen, Germany: 175–180.
- Baek Y., Lee B., Han D., Yun S., Lee H. 2019. Character Region Awareness for Text Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach: 9365–9374.
- Bluche T. 2015. Deep Neural Networks for Large Vocabulary Handwritten Text Recognition. *Computers and Society*, Université Paris Sud, Paris. 20 p.
- Chiang Y.Y., Leyk S., Knoblock C.A. 2014. A Survey of Digital Map Processing Techniques. *ACM Computing Surveys (CSUR)*, 47(1): 1–44. <https://doi.org/10.1145/2557423>
- De Cao N., Izacard G., Riedel S., Petroni F. 2021. Autoregressive Entity Retrieval. *Proceedings of the 9th International Conference on Learning Representations*. Vienna, Austria, ICLR: 1–20.

- Deseilligny M.P., Le Men H., Stamon G. 1995. Character String Recognition on Maps, a Rotation-invariant Recognition Method. *Pattern Recognition Letters*, 16(12): 1297–1310. [https://doi.org/10.1016/0167-8655\(95\)00084-5](https://doi.org/10.1016/0167-8655(95)00084-5)
- Epshtein B., Ofek E., Wexler, Y. 2010. Detecting Text in Natural Scenes with Stroke Width Transform. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, 13–18 June 2010. USA, San Francisco, CA: 2963–2970.
- Fletcher L.A., Kasturi R. 1988. A Robust Algorithm for Text String Separation from Mixed Text/graphics Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6): 910–918.
- Gelbukh A., Levachkine S. 2002. Error Detection and Correction in Toponym Recognition in Cartographic Maps. *IASTED International Conference Geopro-2002*: 1–7.
- Gelbukh A., Levachkine S., Han S.Y. 2003. Resolving Ambiguities in Toponym Recognition in Cartographic Maps. In: *Graphics Recognition. Recent Advances and Perspectives*. Berlin, Heidelberg, Springer Berlin Heidelberg: 75–86.
- Gregory I., Donaldson C., Murrieta-Flores P., Rayson P. 2015. Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *International Journal of Humanities and Arts Computing*, 9(1): 1–14. <https://doi.org/10.3366/ijhac.2015.0135>
- Gritta M., Pilehvar M., Collier N. 2018. Which Melbourne? Augmenting Geocoding with Maps. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Australia, Melbourne: 1285–1296.
- Hu X., Al-Olimat H., Kersten J., Wiegmann M., Klan F., Sun Y., Fan H. 2021. GazPNE: Annotation-free Deep Learning for Place Name Extraction from Microblogs Leveraging Gazetteer and Synthetic Data by Rules. *International Journal of Geographical Information Science*, 36(2): 310–337. <https://doi.org/10.1080/13658816.2021.1947507>
- Hu X., Sun Y., Kersten J., Zhou Z., Klan F., Fan H. 2023. How Can Voting Mechanisms Improve the Robustness and Generalizability of Toponym Disambiguation? *International Journal of Applied Earth Observations and Geoinformation*, 117: 103191. <https://doi.org/10.1016/j.jag.2023.103191>
- Hu X., Zhou Z., Li H., Hu Y., Gu F., Kersten J., Fan H., Klan F. 2022a. Location Reference Recognition from Texts: A Survey and Comparison. *ACM Computing Surveys*, 56(5): 1–37. <https://doi.org/10.1145/3625819>
- Hu X., Zhou Z., Sun Y., Kersten J., Klan F., Fan H., Wiegmann M. 2022b. GazPNE2: A General Place Name Extractor for Microblogs Fusing Gazetteers and Pretrained Transformer Models. *IEEE Internet of Things Journal*, 9(17): 16259–16271. <https://doi.org/10.1109/JIOT.2022.3150967>
- Kent A.J. 2008. Cartographic Blandscapes and the New Noise: Finding the Good View in a Topographical Mashup. *The Bulletin of the Society of Cartographers*, 42(1–2): 29–37.
- Kyamargiou E., Papakondylis Y., Scalora F., Dimitropoulos D. 2020. Changing the Map in Greece and Italy: Place-name Changes in the Nineteenth Century. *The Historical Review/La Revue Historique*, 17: 205–250. <https://doi.org/10.12681/hr.27072>
- Lenc L., Martínek J., Baloun J., Prantl M., Král P. 2022. Historical Map Toponym Extraction for Efficient Information Retrieval. In: *Document Analysis Systems*. Cham: Springer International Publishing: 171–183. https://doi.org/10.1007/978-3-031-06555-2_12
- Milleville K., Verstockt S., Van de Weghe N. 2020. Improving Toponym Recognition Accuracy of Historical Topographic Maps. In: *Automatic Vectorisation of Historical Maps*. International workshop organized by the ICA Commission on Cartographic Heritage into the Digital, Budapest, 13 March 2020. Budapest, Department of Cartography and Geoinformatics: 63–72.
- Neumann L., Matas J. 2010. A Method for Text Localization and Recognition in Real-world Images. In: *Computer Vision – ACCV 2010. Proceedings of the 10th Asian Conference on Computer Vision*, New Zealand, 8–12 November 2010. Berlin, Heidelberg, Springer: 770–783.
- Neumann L., Matas J. 2012. Real-time Scene Text Localization and Recognition. In: *ACM Transactions on Multimedia Computing, Communications, and Applications*. IEEE Conference on Computer Vision and Pattern Recognition, 16–21 June 2012. RI, USA, Providence: 3538–3545.
- Olson R., Kim J., Chiang Y.Y. 2023. An Automatic Approach to Finding Geographic Name Changes on Historical Maps. In: *ACM SIGSPATIAL GIS 2023. Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, Hamburg, 13–16 November 2023. Germany, Hamburg, Association for Computing Machinery: 1–2.



- Olson R., Kim J., Chiang Y.Y. 2024. Automatic Search of Multiword Place Names on Historical Maps. In: GeoSearch 2024. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Searching and Mining Large Collections of Geospatial Data, Atlanta, 29 October 2024. Atlanta, Association for Computing Machinery: 9–12.
- Pezeshk A., Tutwiler R.L. 2011. Automatic Feature Extraction and Text Recognition from Scanned Topographic Maps. *IEEE Transactions on Geoscience and Remote Sensing*, 49(12): 5047–5063. <https://doi.org/10.1109/TGRS.2011.2157697>
- Pierrot-Deseilligny M., Men H.L., Stamon G. 1995. Characters String Recognition on Maps, a Method for High Level Reconstruction. In: ICDAR '95. Proceedings of the Third International Conference on Document Analysis and Recognition, Montreal, 14–16 August 1995. QC, Canada, Montreal: 249–252.
- Pouderoux J., Gonzato J-Ch., Pereira A., Guitton P. 2007. Toponym Recognition in Scanned Color Topographic Maps. In: ICDAR '07. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), Curitiba, 23–26 September 2007. Brazil, Curitiba: 531–535.
- Schlegel I. 2021. Automated Extraction of Labels from Large-scale Historical Maps. *AGILE GIScience Series*, 2(12): 1–14.
- Scott P., Bader M. K-F., Burgess T., Hardy G., Williams N. 2019. Global Biogeography and Invasion Risk of the Plant Pathogen Genus *Phytophthora*. *Environmental Science and Policy*, 101: 175–182. <https://doi.org/10.1016/j.envsci.2019.08.020>
- Sekido I. 2019. Historical GIS Materials for South Asia Studies in the University of Tokyo. *Journal of Urban and Regional Studies on Contemporary India*, 5(2): 23–27.
- Sevgili Ö., Shelmanov A., Arkhipov M., Panchenko A., Biemann C. 2022. Neural Entity Linking: A Survey of Models Based on Deep Learning. *Semantic Web*, 13(3): 527–570. <https://doi.org/10.3233/SW-222986>
- Smith D.A., Crane G.R. 2001. Disambiguating Geographic Names in a Historical Digital Library. In: Research and Advanced Technology for Digital Libraries. International Conference on Theory and Practice of Digital Libraries, Berlin, 30 August 2001. Berlin, Heidelberg, Springer Berlin Heidelberg: 127–136.
- Weissenbacher D., Tahsin T., Beard R., Figaro M., Rivera R., Scotch M., Gonzalez G. 2015. Knowledge-driven Geospatial Location Resolution for Phylogeographic Models of Virus Migration. *Bioinformatics*, 31(12): 348–356. <https://doi.org/10.1093/bioinformatics/btv259>
- Woodruff A.G., Plaunt Chr. 1994. GIPSY: Automated Geographic Indexing of Text Documents. *Journal of the American Society for Information Science*, 45(9): 645–655.
- Wu L., Petroni F., Josifoski M., Riedel S., Zettlemoyer L. 2020. Zero-shot Entity Linking with Dense Entity Retrieval. In: Computation and Language. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, 16–20 November 2020. Stroudsburg, Association for Computational Linguistics: 6397–6407.
- Zhang Y., Chen Z., Zheng X., Chen N., Wang Y. 2021. Extracting the Location of Flooding Events in Urban Systems and Analyzing the Semantic Risk Using Social Sensing Data. *Journal of Hydrology*, 603(4): 127053. <https://doi.org/10.1016/j.jhydrol.2021.127053>
- Zhou X., Yao C., Wen H., Wang Y., Zhou S., He W., Liang J. 2017. EAST: an Efficient and Accurate Scene Text Detector. In: Computer Science. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 21–26 July 2017. HI, USA, Honolulu: 5551–5560.

References

- Analakova O., Yollyyeva O. 2024. Using Artificial Intelligence to Analyze Spatial Data and Identify Patterns in Cartography. *Innovacionnaja nauka*, 5–2–2: 158–160 (in Russian).
- Batalov D.S., Bessmertniy I.V. 2023. *Primenenie iskusstvennogo intellekta v toponimike. Obzor metodov iskusstvennogo intellekta i nauchnyh rabot* [Application of Artificial Intelligence in Toponymy. Review of Artificial Intelligence Methods and Scientific Works]. In: *Geograficheskie aspekty ustojchivogo razvitiya regionov* [Geographical Aspects of Sustainable Development of Regions]. V International Scientific and Practical Conference, Gomel, 25–26 May 2023. Gomel, Publ. Gomelskiy gosudarstvennyy universitet imeni Frantsiska Skoriny: 372–376.

- Boyarsky K.K., Kanevsky E.A., Butorina D.A. 2019. On the Issue of the Semantic Disambiguation of Toponyms in Russian Texts. *Computer Linguistics and Computing Ontologies*, 3: 19–28 (in Russian). <https://doi.org/10.17586/2541-9781-2019-3-19-28>
- Vicentiy A.V., Dikovitsky V.V., Shishaev M.G. 2020. The Technology of Extraction and Visualization of Spatial Data Obtained by Texts Analysis. *Kola Science Centre Publisher*, 11(8(11)): 115–119 (in Russian). <https://doi.org/10.37614/2307-5252.2020.8.11.012>
- Vicentiy A.V., Shishaev M.G. 2021. The Geoattributed Entity Extraction Technology for Visual Representation of Objects Spatial Relations Based on Automated Schematic Map Generation. *Transactions of the Kola Science Centre. Information technologies*, 12(5): 35–49 (in Russian)
- Dmitriev A.V. 2025. The Early Stage in the Methodology of Automatic Toponym Recognition on Geographic Maps and in Texts. *Regional Geosystems*, 49(1): 128–145 (in Russian). <https://doi.org/10.52575/2712-7443-2025-49-1-128-145>
- Kreydlin L.G. 2005. TagLite: The Program of Identification of Russian Individualized NPs. In: *Computational Linguistics and Intellectual Technologies. Annual International Conference "Dialogue"*, Moscow, 01–04 June 2016. Moscow, Publ. Nauka: 292–297 (in Russian).
- Ardanuy M.C., Sporleder C. 2017. Toponym Disambiguation in Historical Documents Using Semantic and Geographic Features. *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage*, Göttingen, Germany: 175–180.
- Baek Y., Lee B., Han D., Yun S., Lee H. 2019. Character Region Awareness for Text Detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach: 9365–9374.
- Bluche T. 2015. Deep Neural Networks for Large Vocabulary Handwritten Text Recognition. *Computers and Society*, Université Paris Sud, Paris. 20 p.
- Chiang Y.Y., Leyk S., Knoblock C.A. 2014. A Survey of Digital Map Processing Techniques. *ACM Computing Surveys (CSUR)*, 47(1): 1–44. <https://doi.org/10.1145/2557423>
- De Cao N., Izacard G., Riedel S., Petroni F. 2021. Autoregressive Entity Retrieval. *Proceedings of the 9th International Conference on Learning Representations*. Vienna, Austria, ICLR: 1–20.
- Deseilligny M.P., Le Men H., Stamon G. 1995. Character String Recognition on Maps, a Rotation-invariant Recognition Method. *Pattern Recognition Letters*, 16(12): 1297–1310. [https://doi.org/10.1016/0167-8655\(95\)00084-5](https://doi.org/10.1016/0167-8655(95)00084-5)
- Epshtein B., Ofek E., Wexler, Y. 2010. Detecting Text in Natural Scenes with Stroke Width Transform. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Francisco, 13–18 June 2010. USA, San Francisco, CA: 2963–2970.
- Fletcher L.A., Kasturi R. 1988. A Robust Algorithm for Text String Separation from Mixed Text/graphics Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10(6): 910–918.
- Gelbukh A., Levachkine S. 2002. Error Detection and Correction in Toponym Recognition in Cartographic Maps. *IASTED International Conference Geopro-2002*: 1–7.
- Gelbukh A., Levachkine S., Han S.Y. 2003. Resolving Ambiguities in Toponym Recognition in Cartographic Maps. In: *Graphics Recognition. Recent Advances and Perspectives*. Berlin, Heidelberg, Springer Berlin Heidelberg: 75–86.
- Gregory I., Donaldson C., Murrieta-Flores P., Rayson P. 2015. Geoparsing, GIS, and Textual Analysis: Current Developments in Spatial Humanities Research. *International Journal of Humanities and Arts Computing*, 9(1): 1–14. <https://doi.org/10.3366/ijhac.2015.0135>
- Gritta M., Pilehvar M., Collier N. 2018. Which Melbourne? Augmenting Geocoding with Maps. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Australia, Melbourne: 1285–1296.
- Hu X., Al-Olimat H., Kersten J., Wiegmann M., Klan F., Sun Y., Fan H. 2021. GazPNE: Annotation-free Deep Learning for Place Name Extraction from Microblogs Leveraging Gazetteer and Synthetic Data by Rules. *International Journal of Geographical Information Science*, 36(2): 310–337. <https://doi.org/10.1080/13658816.2021.1947507>
- Hu X., Sun Y., Kersten J., Zhou Z., Klan F., Fan H. 2023. How Can Voting Mechanisms Improve the Robustness and Generalizability of Toponym Disambiguation? *International Journal of Applied Earth Observations and Geoinformation*, 117: 103191. <https://doi.org/10.1016/j.jag.2023.103191>
- Hu X., Zhou Z., Li H., Hu Y., Gu F., Kersten J., Fan H., Klan F. 2022a. Location Reference Recognition from Texts: A Survey and Comparison. *ACM Computing Surveys*, 56(5): 1–37. <https://doi.org/10.1145/3625819>



- Hu X., Zhou Z., Sun Y., Kersten J., Klan, F., Fan H., Wiegmann M. 2022b. GazPNE2: A General Place Name Extractor for Microblogs Fusing Gazetteers and Pretrained Transformer Models. *IEEE Internet of Things Journal*, 9(17): 16259–16271. <https://doi.org/10.1109/JIOT.2022.3150967>
- Kent A.J. 2008. Cartographic Blandscapes and the New Noise: Finding the Good View in a Topographical Mashup. *The Bulletin of the Society of Cartographers*, 42(1–2): 29–37.
- Kyramargiou E., Papakondylis Y., Scalora F., Dimitropoulos D. 2020. Changing the Map in Greece and Italy: Place-name Changes in the Nineteenth Century. *The Historical Review/La Revue Historique*, 17: 205–250. <https://doi.org/10.12681/hr.27072>
- Lenc L., Martínek J., Baloun J., Prantl M., Král P. 2022. Historical Map Toponym Extraction for Efficient Information Retrieval. In: *Document Analysis Systems*. Cham: Springer International Publishing: 171–183. https://doi.org/10.1007/978-3-031-06555-2_12
- Milleville K., Verstockt S., Van de Weghe N. 2020. Improving Toponym Recognition Accuracy of Historical Topographic Maps. In: *Automatic Vectorisation of Historical Maps*. International workshop organized by the ICA Commission on Cartographic Heritage into the Digital, Budapest, 13 March 2020. Budapest, Department of Cartography and Geoinformatics: 63–72.
- Neumann L., Matas J. 2010. A Method for Text Localization and Recognition in Real-world Images. In: *Computer Vision – ACCV 2010. Proceedings of the 10th Asian Conference on Computer Vision*, New Zealand, 8–12 November 2010. Berlin, Heidelberg, Springer: 770–783.
- Neumann L., Matas J. 2012. Real-time Scene Text Localization and Recognition. In: *ACM Transactions on Multimedia Computing, Communications, and Applications. IEEE Conference on Computer Vision and Pattern Recognition*, 16–21 June 2012. RI, USA, Providence: 3538–3545.
- Olson R., Kim J., Chiang Y.Y. 2023. An Automatic Approach to Finding Geographic Name Changes on Historical Maps. In: *ACM SIGSPATIAL GIS 2023. Proceedings of the 31st ACM International Conference on Advances in Geographic Information Systems*, Hamburg, 13–16 November 2023. Germany, Hamburg, Association for Computing Machinery: 1–2.
- Olson R., Kim J., Chiang Y.Y. 2024. Automatic Search of Multiword Place Names on Historical Maps. In: *GeoSearch 2024. Proceedings of the 3rd ACM SIGSPATIAL International Workshop on Searching and Mining Large Collections of Geospatial Data*, Atlanta, 29 October 2024. Atlanta, Association for Computing Machinery: 9–12.
- Pezeshk A., Tutwiler R.L. 2011. Automatic Feature Extraction and Text Recognition from Scanned Topographic Maps. *IEEE Transactions on Geoscience and Remote Sensing*, 49(12): 5047–5063. <https://doi.org/10.1109/TGRS.2011.2157697>
- Pierrot-Deseilligny M., Men H.L., Stamon G. 1995. Characters String Recognition on Maps, a Method for High Level Reconstruction. In: *ICDAR '95. Proceedings of the Third International Conference on Document Analysis and Recognition*, Montreal, 14–16 August 1995. QC, Canada, Montreal: 249–252.
- Pouderoux J., Gonzato J-Ch., Pereira A., Guitton P. 2007. Toponym Recognition in Scanned Color Topographic Maps. In: *ICDAR '07. Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, Curitiba, 23–26 September 2007. Brazil, Curitiba: 531–535.
- Schlegel I. 2021. Automated Extraction of Labels from Large-scale Historical Maps. *AGILE GIScience Series*, 2(12): 1–14.
- Scott P., Bader M. K-F., Burgess T., Hardy G., Williams N. 2019. Global Biogeography and Invasion Risk of the Plant Pathogen Genus *Phytophthora*. *Environmental Science and Policy*, 101: 175–182. <https://doi.org/10.1016/j.envsci.2019.08.020>
- Sekido I. 2019. Historical GIS Materials for South Asia Studies in the University of Tokyo. *Journal of Urban and Regional Studies on Contemporary India*, 5(2): 23–27.
- Sevgili Ö., Shelmanov A., Arkhipov M., Panchenko A., Biemann C. 2022. Neural Entity Linking: A Survey of Models Based on Deep Learning. *Semantic Web*, 13(3): 527–570. <https://doi.org/10.3233/SW-222986>
- Smith D.A., Crane G.R. 2001. Disambiguating Geographic Names in a Historical Digital Library. In: *Research and Advanced Technology for Digital Libraries. International Conference on Theory and Practice of Digital Libraries*, Berlin, 30 August 2001. Berlin, Heidelberg, Springer Berlin Heidelberg: 127–136.
- Weissenbacher D., Tahsin T., Beard R., Figaro M., Rivera R., Scotch M., Gonzalez G. 2015. Knowledge-driven Geospatial Location Resolution for Phylogeographic Models of Virus Migration. *Bioinformatics*, 31(12): 348–356. <https://doi.org/10.1093/bioinformatics/btv259>

- Woodruff A.G., Plaunt Chr. 1994. GIPSY: Automated Geographic Indexing of Text Documents. Journal of the American Society for Information Science, 45(9): 645–655.
- Wu L., Petroni F., Josifoski M., Riedel S., Zettlemoyer L. 2020. Zero-shot Entity Linking with Dense Entity Retrieval. In: Computation and Language. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, 16–20 November 2020. Stroudsburg, Association for Computational Linguistics: 6397–6407.
- Zhang Y., Chen Z., Zheng X., Chen N., Wang Y. 2021. Extracting the Location of Flooding Events in Urban Systems and Analyzing the Semantic Risk Using Social Sensing Data. Journal of Hydrology, 603(4): 127053. <https://doi.org/10.1016/j.jhydrol.2021.127053>
- Zhou X., Yao C., Wen H., Wang Y., Zhou S., He W., Liang J. 2017. EAST: an Efficient and Accurate Scene Text Detector. In: Computer Science. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 21–26 July 2017. HI, USA, Honolulu: 5551–5560.

*Поступила в редакцию 23.03.2025;
поступила после рецензирования 06.05.2025;
принята к публикации 12.06.2025*

*Received March 23, 2025;
Revised May 06, 2025;
Accepted June 12, 2025*

Конфликт интересов: о потенциальном конфликте интересов не сообщалось.

Conflict of interest: no potential conflict of interest related to this article was reported.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Крупнова Елена Сергеевна, специалист по учебно-методической работе 1 категории учебно-научной Лаборатории цифровой лингвистики, Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Россия

Дмитриев Александр Владиславович, кандидат филологических наук, доцент Высшей школы лингвистики и педагогики, заведующий учебно-научной Лабораторией цифровой лингвистики, Санкт-Петербургский политехнический университет Петра Великого, г. Санкт-Петербург, Россия

INFORMATION ABOUT THE AUTHORS

Elena S. Krupnova, specialist in educational and methodological work of first category in the Educational and Scientific Laboratory of Digital Linguistics, Peter the Great St. Petersburg Polytechnic University, Saint-Petersburg, Russia

Alexander V. Dmitriev, Candidate of Philological Sciences, Associate Professor at the Higher School of Linguistics and Pedagogy of Peter the Great St. Petersburg Polytechnic University, Head of the Educational and Scientific Laboratory of Digital Linguistics, Saint-Petersburg, Russia