

А.Р. Гири́н, Н.Н. Тесля
**МЕТОД СОПОСТАВЛЕНИЯ ЗАПИСЕЙ О ДТП
ИЗ ОФИЦИАЛЬНЫХ ИСТОЧНИКОВ
И НОВОСТНЫХ ПУБЛИКАЦИЙ НА ОСНОВЕ
ПРОСТРАНСТВЕННО-ВРЕМЕННОЙ КЛАСТЕРИЗАЦИИ**

Гирин А.Р., Тесля Н.Н. Метод сопоставления записей о ДТП из официальных источников и новостных публикаций на основе пространственно-временной кластеризации.

Аннотация. Принятие эффективных решений для снижения уровня аварийности на автомобильном транспорте основывается на результатах анализа мест концентрации ДТП. Полнота и точность данных о ДТП, используемых для анализа, могут быть повышены за счет интеграции нескольких источников. Важным этапом интеграции является сопоставление записей из различных источников, описывающих одни и те же события ДТП. Решение задачи сопоставления осложняется проблемой гетерогенности: записи о ДТП могут отличаться по структурированности и формату представления. В исследовании предложен метод решения задачи сопоставления записей о ДТП из двух гетерогенных источников: официальных ресурсов со структурированными карточками учета ДТП и новостных публикаций в виде неструктурированных текстов на естественном языке. Предложенный метод основан на гипотезе о пространственной и временной близости: записи из различных источников считаются соответствующими одному ДТП, если описанные в них события произошли на небольшом расстоянии и с незначительной временной разницей. Метод включает два последовательных этапа. На первом этапе из новостных публикаций извлекаются координаты и дата/время описываемого события с использованием экспертных правил, географического справочника и большой языковой модели LLaMA. На втором этапе выполняется сопоставление записей из официальных ресурсов и новостных публикаций на основе пространственно-временной кластеризации с применением алгоритма ST-DBSCAN. Метод позволяет дополнять описание контекста официально зарегистрированных ДТП и обогащать доступные для анализа данные за счет включения ДТП, сведения о которых отсутствуют в официальных источниках. Экспериментальная оценка метода проведена на реальных данных о ДТП в Санкт-Петербурге. Результаты показали, что комбинирование классических методов с большой языковой моделью повышает точность извлечения пространственно-временных атрибутов ДТП из новостных публикаций, а применение пространственно-временной кластеризации обеспечивает высокую точность сопоставления записей о ДТП (F1-мера = 0.84).

Ключевые слова: интеграция данных, сопоставление записей, дорожно-транспортные происшествия, газетир, экспертные правила, ST-DBSCAN, большие языковые модели, LLaMA.

1. Введение. Дорожно-транспортные происшествия (ДТП) представляют собой значительную социально-экономическую проблему, актуальную как на мировом уровне, так и в России. Согласно официальным данным, в 2024 году в стране произошло более 132 тыс. ДТП, в результате которых погибло свыше 14 тыс. человек и более 164 тыс. получили травмы [1]. Принятие эффективных управленческих решений, направленных на снижение уровня аварийности, основывается

на результатах анализа мест концентрации ДТП и выявлении причин их возникновения. Качество такого анализа и, как следствие, эффективность принимаемых решений напрямую зависят от полноты и точности доступных исследователям данных о ДТП [2].

В международной практике повышение полноты и точности данных о ДТП достигается за счет интеграции нескольких источников. Записи из различных источников, описывающие одни и те же события ДТП, сопоставляются между собой с последующим объединением. Это позволяет уточнять и обогащать описание контекста ДТП, а также включать в анализ данные о ранее не учтенных ДТП [3].

Как правило, в исследованиях основным источником данных о ДТП выступают записи из официальных ресурсов органов государственной власти. К числу дополнительных источников относятся данные бортовых устройств, записи устройств дорожной инфраструктуры, а также новостные публикации в СМИ и социальных сетях [4]. Новостные публикации являются особенно перспективным источником. Их распространенность и общедоступность обеспечивают широкий охват ДТП, в том числе событий, сведения о которых не публикуются на официальных ресурсах (например, ДТП без пострадавших). Кроме того, публикации содержат детализированное описание контекста ДТП, включая данные о местоположении, времени, участниках и возможных причинах ДТП, что представляет ценность для анализа [5, 6].

Интеграция записей из официальных ресурсов и новостных публикаций позволяет повысить качество данных о ДТП, используемых для анализа [7]. Ключевым этапом интеграции является решение задачи сопоставления записей из этих источников (*record linkage*), которая состоит в определении записей, описывающих одно и то же событие ДТП. Такое сопоставление осложняется проблемой гетерогенности источников. Если записи из официальных ресурсов имеют строгую структуру и регламентированный набор атрибутов, то новостные публикации представляют собой неструктурированный текст на естественном языке, что, наряду с проблемой отсутствия между источниками единого сквозного идентификатора описываемых событий ДТП, препятствует прямому сопоставлению записей [8].

Существующие методы решения задачи сопоставления записей могут быть разделены на несколько групп. К первой группе относятся классические детерминированные и вероятностные методы (*probabilistic record linkage*) [9, 10]. Детерминированные методы используют строгие правила или функции сходства строк (например, расстояние Левенштейна или Джаро-Винклера)

для сравнения атрибутов. Вероятностные методы, основанные на модели Феллелджи-Сантера, оценивают вероятность совпадения записей на основе весов согласия и расхождения признаков. Основным преимуществом данных методов является их интерпретируемость и вычислительная эффективность, однако возможность их применения ограничена необходимостью наличия строго структурированных наборов атрибутов. Адаптация данных методов за счет предварительного извлечения полного набора атрибутов описываемого в тексте события ДТП сопряжена с риском накопления ошибок [11].

Отдельно может быть выделен метод сопоставления записей о ДТП на основе гипотезы о пространственной и временной близости: записи считаются соответствующими одному событию ДТП, если описанные в них события произошли на небольшом расстоянии друг от друга и с незначительной временной разницей. Гипотеза использована, например, при сопоставлении записей о ДТП из официального ресурса дорожной полиции и социального навигационного приложения Waze в городе Белу-Оризонти (Бразилия) [4]. Получение значений пространственно-временных атрибутов ДТП в данном исследовании осуществлялось на основе метаданных записей (геометок). Использование метаданных позволило исключить этап обработки текста публикации для извлечения атрибутов, однако применимость данного метода к решению поставленной задачи ограничена, так как метаданными сопровождается лишь малая часть новостных публикаций о ДТП в СМИ и социальных сетях.

Вторую группу составляют методы на основе классического машинного обучения, где задача сопоставления формулируется как задача бинарной классификации пар записей [12]. Несмотря на их способность выявлять более сложные зависимости между признаками по сравнению с детерминированными методами, их использование в предметной области исследования ограничено. Это обусловлено отсутствием обширных размеченных наборов данных с записями о ДТП, необходимых для обучения моделей, а также сложностью формирования репрезентативного вектора признаков из неструктурированного текста публикаций.

Третья группа включает методы глубокого обучения на основе предобученных языковых моделей, таких как BERT или RoBERTa, и специализированных систем, например, Ditto [13, 14]. Использование данных методов позволяет достигать высоких показателей F1-меры и обеспечивать устойчивость к пропускам или ошибкам в атрибутах. Однако высокая сложность внутренней архитектуры моделей затрудняет

интерпретацию логики принятия решений. Кроме того, данные методы требуют значительных объемов обучающих данных для тонкой настройки (fine-tuning) и часто показывают низкую точность на данных, не представленных в обучающей выборке [15].

Четвертую группу представляют методы, использующие генеративные большие языковые модели (LLMs), такие как GPT-4. Эти методы позволяют решать задачу сопоставления в режимах zero-shot (без примеров) или few-shot (с минимальным количеством примеров) за счет использования контекстного обучения и промпт-инжиниринга [16, 17]. Исследования показывают, что LLM способны генерировать структурированные объяснения принятых решений и обладают высокой способностью к обобщению на новых типах данных [18]. Тем не менее, методы на основе LLM характеризуются высокой чувствительностью к формулировкам запросов, а их применение может быть ограничено стоимостью API-запросов или высокими требованиями к вычислительным ресурсам при локальном развертывании моделей [19]. Кроме того, применение LLM может сопровождаться галлюцинациями модели.

Обзор существующих методов сопоставления записей позволяет сделать вывод о наличии значимых ограничений при их работе с неструктурированными текстами новостных публикаций. Детерминированные и вероятностные методы эффективны преимущественно при работе со структурированными данными. Использование классического машинного и глубокого обучения требует подготовки обширных размеченных наборов данных о ДТП, доступность которых в предметной области ограничена. Применение больших языковых моделей обеспечивает высокую точность сопоставления записей без предварительной настройки, однако такие модели чувствительны к формулировкам запросов и характеризуются высокими вычислительными затратами и возможными галлюцинациями модели.

Целью исследования является разработка метода сопоставления записей о ДТП из официальных ресурсов и новостных публикаций, позволяющего преодолеть ограничения существующих методов. В основу исследования положены следующие гипотезы:

1. Сопоставление записей о ДТП из разнородных источников может быть выполнено на основании сравнения значений их пространственно-временных атрибутов и проверки условия пространственной и временной близости описываемых в записях событий.

2. Пространственно-временные атрибуты могут быть извлечены непосредственно из текстов новостных публикаций, при этом точность

извлечения повышается за счет использования комбинированного метода: комбинации классических методов и большой языковой модели, способной интерпретировать сложные языковые конструкции.

На основе выдвинутых гипотез в исследовании предлагается двухэтапный метод сопоставления записей. На первом этапе из текстов новостных публикаций извлекаются координаты и дата/время описываемого события ДТП с использованием разработанной процедуры, объединяющей экспертные правила, географический справочник (газетир) и большую языковую модель LLaMA [20]. На втором этапе записи из официальных ресурсов и новостных публикаций сопоставляются на основе пространственно-временной кластеризации с использованием алгоритма ST-DBSCAN [21].

Описание и экспериментальная оценка метода выполнены на примере источников данных о ДТП в Санкт-Петербурге. В качестве официального ресурса использован сайт Госавтоинспекции [1], на котором размещаются карточки учета официально зарегистрированных ДТП. Источником новостных публикаций выступили региональные интернет-СМИ и тематические сообщества в социальной сети «ВКонтакте». На текущем этапе исследования рассмотрена исключительно текстовая часть новостных публикаций на русском языке, а в анализ включены только те записи, которые содержат описание единственного события ДТП. При этом предлагаемый метод не ограничен спецификой конкретных источников и может быть адаптирован для использования в других регионах.

Научная новизна предлагаемого подхода заключается в разработке гибридного метода сопоставления записей, основанного на концепции двухэтапной верификации пространственно-временных атрибутов. В отличие от существующих подходов, ориентированных на полное извлечение контекста ДТП, предложенный метод базируется на гипотезе о достаточности пространственно-временной близости записей для идентификации уникального события.

В отличие от современных (state-of-the-art) систем сопоставления (например, Ditto [13]), требующих наличия тысяч размеченных пар записей для обучения, предлагаемый метод ориентирован на работу в условиях отсутствия обучающих выборок. Прямое сравнение показывает, что использование LLM в режиме верификации в комбинации с классическими методами позволяет достичь точности сопоставления ($F1$ -мера = 0.84), сопоставимой с результатами специализированных трансформеров на эталонных наборах данных [18],

при существенном снижении трудозатрат на предварительную подготовку и разметку данных о ДТП.

2. Формализация предметной области исследования. В рамках исследования решается задача сопоставления записей о ДТП, полученных из официальных карточек учета ДТП и новостных публикаций в СМИ и социальных сетях. Задача сопоставления заключается в определении записей, описывающих одно и то же событие ДТП, и формально может быть поставлена в терминах классической задачи разрешения сущностей (Entity Resolution) [10].

Под сущностью понимается объект или событие реального мира, а разрешение сущностей заключается в сопоставлении записей из различных источников данных, описывающих одни и те же сущности. В предметной области исследования основной сущностью выступает ДТП. Источниками данных, содержащими описания ДТП, являются официальные карточки учета ДТП от Госавтоинспекции и новостные публикации в СМИ и социальных сетях. Для формальной постановки задачи сопоставления и описания предлагаемого метода ее решения вводится ряд обозначений.

2.1. Используемые обозначения. ДТП рассматривается как сложная сущность, состоящая из нескольких взаимосвязанных элементов: участников ДТП, транспортных средств и внешних условий (погодные, дорожные и другие условия). Состояние этих элементов в момент ДТП определяет контекст ДТП, который может быть представлен в виде множества пар «атрибут–значение»:

- Атрибуты участников ДТП (роль в ДТП, возраст, стаж вождения, состояние здоровья, наличие признаков опьянения, использование средств пассивной безопасности и другие);

- Атрибуты транспортных средств (категория, марка и модель, год выпуска, наличие технических неисправностей, характер и локализация повреждений и другие);

- Атрибуты внешней среды (погодные условия, тип и состояние дорожного покрытия, уровень естественной и искусственной освещенности, параметры транспортного потока, обустройство участка дороги и другие).

Пусть через E обозначается множество ДТП, произошедших в исследуемом регионе за фиксированный промежуток времени; через A множество атрибутов, используемых для описания контекста ДТП; через R множество доступных в исследовании записей, содержащих сведения о ДТП из множества E (примерами записи выступают официальные карточки учета ДТП и новостные публикации). В соответствии

с введенными ограничениями исследования, каждая запись содержит описание единственного события ДТП.

Поскольку формат записей может быть разным (от структурированных карточек учета ДТП до неструктурированного текста в новостных публикациях), значения атрибутов не всегда могут быть получены напрямую и требуют извлечения из записей. Пусть через $a(r)$ обозначается значение атрибута $a \in A$, извлеченное из записи $r \in R$. Для любой записи $r \in R$ через $e(r) \in E$ обозначается ДТП, сведения о котором содержатся в данной записи.

Записи $r_1, r_2 \in R$ считаются эквивалентными, если они описывают одно и то же событие ДТП, то есть для них выполняется условие $e(r_1) = e(r_2)$. Каждое реальное событие ДТП $e \in E$ характеризуется вектором параметров $\Theta_e = \langle L_e, T_e \rangle$, где L_e - точные географические координаты, а T_e - точное время события. Критерием истинного совпадения двух записей $r_1, r_2 \in R$ является равенство векторов их параметров:

$$e(r_1) = e(r_2) \iff \Theta_{e(r_1)} = \Theta_{e(r_2)}. \quad (1)$$

2.2. Описание источников данных о ДТП. Официальным источником данных о ДТП в России является сайт Госавтоинспекции, на котором публикуются карточки учета ДТП. Пусть через $R_{official} \subseteq R$ обозначается множество карточек учета ДТП для исследуемого региона за фиксированный промежуток времени. Структура карточек учета строго регламентирована и каждое официально зарегистрированное ДТП описывается только одной карточкой учета:

$$\forall r_1, r_2 \in R_{official}, e(r_1) = e(r_2) \implies r_1 = r_2. \quad (2)$$

В качестве альтернативного источника данных о ДТП рассматриваются новостные публикации в СМИ и социальных сетях. Пусть через $R_{news} \subseteq R$ обозначается множество записей о ДТП в исследуемом регионе за фиксированный промежуток времени, полученных из этих источников. В соответствии с введенными ограничениями исследования, каждая запись рассматривается исключительно как текстовое содержимое публикации. В отличие от карточек учета ДТП, одно и то же ДТП может описываться в нескольких публикациях. Для записей $r \in R_{news}$ определение

значений атрибутов $a(r)$ требует применения методов обработки естественного языка.

2.3. Формальная постановка задачи сопоставления записей.

С учетом введенных обозначений, задача сопоставления записей о ДТП, полученных из официальных карточек учета ДТП и новостных публикаций, формально ставится следующим образом.

Пусть $R = R_{official} \cup R_{news}$ множество всех доступных в исследовании записей о ДТП. Необходимо построить разбиение множества R на кластеры:

$$\mathcal{P} = \{R_1, \dots, R_n\}, \quad (3)$$

для которых выполняются следующие условия:

1. $\forall i \neq j \quad R_i \cap R_j = \emptyset$.
2. $\bigcup_{i=1}^n R_i = R$.
3. $\forall i \forall r_1, r_2 \in R_i \quad e(r_1) = e(r_2)$.
4. $\forall r_1, r_2 \in R \quad e(r_1) = e(r_2) \implies \exists k : r_1, r_2 \in R_k$.

$R = R_{official} \cup R_{news}$ – доступные описания событий; $e(r) \in E$ – событие, данные о котором описывает $r \in R$; $r_1, r_2 \in R$ считаются эквивалентными, если $e(r_1) = e(r_2)$.

Искомое разбиение $\mathcal{P}$ представляет собой фактор-множество множества R по отношению эквивалентности записей, где каждый кластер является классом эквивалентности, объединяющим все записи об одном уникальном событии ДТП.

В каждом кластере записей R_i может содержаться не более одной записи из $R_{official}$, поскольку в каждом кластере описывается единственное ДТП, а каждое ДТП может быть представлено только в одной карточке учета ДТП. Таким образом, искомое разбиение множества R будет состоять из кластеров двух типов:

1. Кластеры, содержащие одну запись из $R_{official}$ и ноль или более записей из R_{news} . Такие кластеры соответствуют ДТП, сведения о которых были официально опубликованы.

2. Кластеры, содержащие исключительно записи из R_{news} . Такие кластеры соответствуют ДТП, сведения о которых не были официально опубликованы, но были описаны в СМИ и социальных сетях.

3. Предлагаемый метод сопоставления записей. В рамках исследования решается задача сопоставления записей о ДТП. Задача формально поставлена как разбиение множества записей R на непересекающиеся кластеры, каждый из которых соответствует одному

уникальному ДТП. Построение такого разбиения требует возможности для любой пары записей проверить условие их эквивалентности (т.е. соответствия одному ДТП). На практике, однако, эта проверка не может быть выполнена напрямую в силу гетерогенности источников. Так как официальные карточки и новостные публикации формируются независимо, между ними отсутствует общий сквозной идентификатор, позволяющий однозначно установить эквивалентность записей.

Для преодоления этого ограничения предлагается метод сопоставления записей на основе гипотезы, согласно которой две записи считаются эквивалентными, если описанные в них события ДТП произошли на небольшом расстоянии друг от друга и с незначительной временной разницей (Рис. 1). В связи с этим из множества A выделяется ключевое подмножество атрибутов $a_{loc_coords}, a_{datetime} \in A$ (координаты и дата/время ДТП соответственно), совокупность значений которых является достаточной для условно уникальной идентификации ДТП.

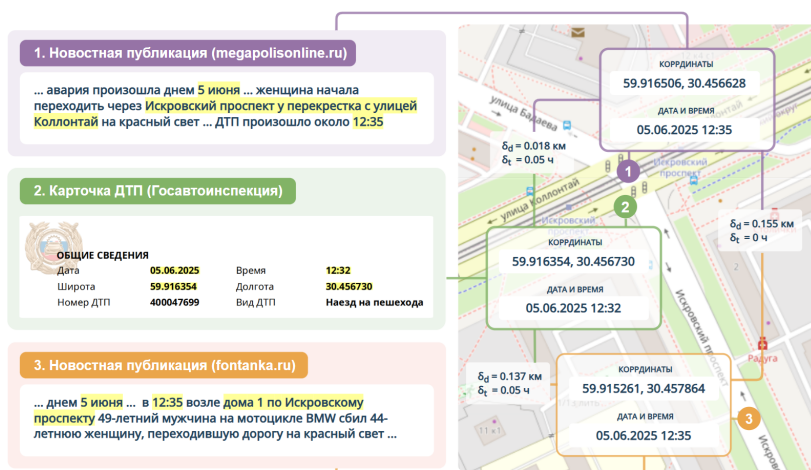


Рис. 1. Демонстрация близости значений пространственно-временных атрибутов для трех записей с описанием одного события ДТП

Практическое решение задачи сопоставления записей о ДТП сводится к последовательному решению следующих подзадач:

1. Разработка процедуры извлечения пространственно-временных атрибутов (a_{loc_coords} и $a_{datetime}$) из записей R .

2. Построение функции схожести записей, которая оценивает соответствие двух записей одному событию ДТП на основе сравнения значений их пространственно-временных атрибутов.

3. Разработка алгоритма сопоставления записей, который на основе построенной функции схожести формирует искомое разбиение множества записей R на кластеры.

Таким образом, предлагаемый метод сопоставления записей о ДТП заключается в последовательном выполнении двух этапов: извлечение пространственно-временных атрибутов из записей R с помощью разработанной процедуры и применение алгоритма сопоставления записей (Рис. 2).

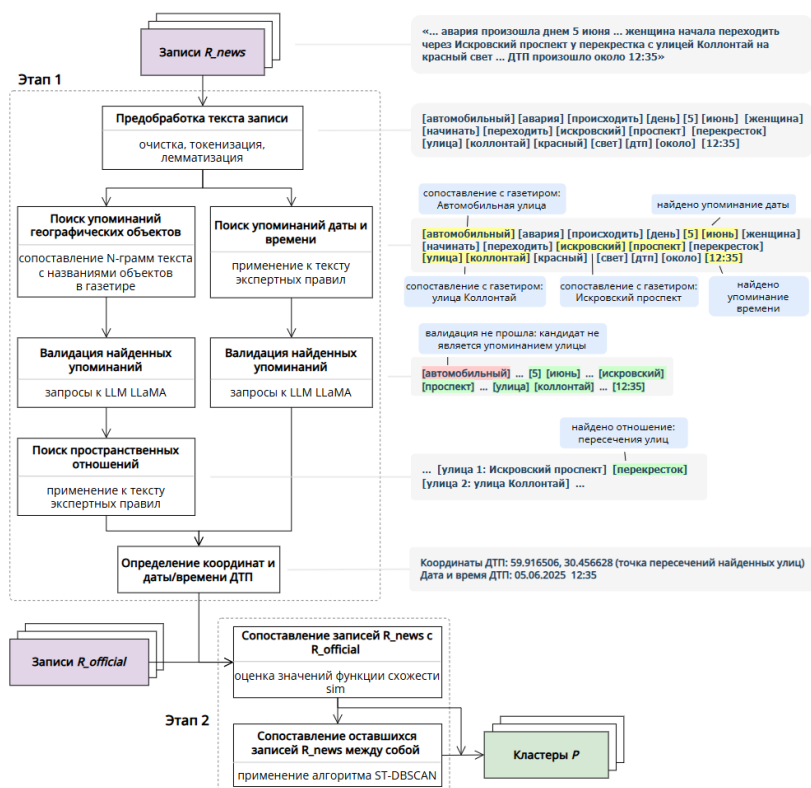


Рис. 2. Общая схема сопоставления записей о ДТП предлагаемым методом

3.1. Процедура извлечения пространственно-временных атрибутов из записей. Для записей из $R_{official}$ значения пространственно-временных атрибутов доступны напрямую из структурированных полей карточек учета ДТП. С другой стороны, определение значений атрибутов для записей R_{news} является нетривиальной задачей. Это обусловлено тем, что новостные публикации, как правило, не сопровождаются метаданными (например, геометками), поэтому значения атрибутов необходимо извлекать непосредственно из неструктурированного текста на русском языке.

Для извлечения пространственно-временных атрибутов из записей R_{news} предлагается процедура, основанная на совместном использовании классических методов (правил и географических справочников) для первичного поиска потенциальных упоминаний в тексте даты/времени и географических объектов, а также большой языковой модели LLM LLaMA для последующей валидации этих упоминаний. Валидация включает две проверки: исключение ложных срабатываний и определение релевантности упоминаний, то есть установление того, относится ли найденная дата/время или географический объект непосредственно к событию ДТП.

Применение методов на основе правил на этапе первичного поиска обеспечивает высокую прозрачность и интерпретируемость логики извлечения кандидатов (потенциальных упоминаний в тексте даты/времени и географических объектов) [22, 23]. Для локализации географических объектов в пределах ограниченной области, в которой названия объектов преимущественно не повторяются, эффективно использование географического справочника (газетира) [24, 25]. В то же время от использования методов классического машинного обучения [26, 27] и глубоких нейронных сетей [14] было решено отказаться ввиду отсутствия обширных размеченных корпусов данных и высоких трудозатрат на их аннотирование. Применение большой языковой модели LLM LLaMA мотивировано ее способностью учитывать семантический контекст всей публикации и интерпретировать сложные языковые конструкции. Это позволяет компенсировать ограничения классических методов [28].

Разработанная процедура включает два блока, отвечающих соответственно за извлечение даты/времени и местоположения события ДТП.

3.1.1. Извлечение даты и времени события ДТП. Извлечение даты и времени события ДТП из текста новостной публикации проводится в три последовательных этапа:

1. Предобработка текста. Входной текст приводится к стандартизованному формату, удобному для дальнейшего анализа. Этот этап включает очистку текста от нерелевантных символов, удаление лишних пробелов, приведение к нижнему регистру, разбиение на отдельные лексемы (токенизацию), приведение слов к их начальной форме (лемматизацию).

2. Применение правил. После предобработки к тексту применяется набор заранее составленных экспертных правил для поиска упоминаний даты и времени. Правила разработаны на основе анализа форматов представления даты и времени в русскоязычных текстах и охватывают как точные (например, «12:30»), так и приблизительные упоминания времени (например, «около часа дня»), а также абсолютные (например, «12 мая 2024 г.») и относительные (например, «вчера») форматы дат.

3. Валидация. Кандидаты (все упоминания даты и времени, найденные на предыдущем этапе) проходят валидацию. Для каждого кандидата последовательно проводятся две проверки с использованием специально разработанных запросов к LLM LLaMA:

- проверка корректности извлечения. Определяется, действительно ли кандидат (например, «16:40») соответствует какому-либо упоминанию даты и времени в исходном тексте, даже если оно представлено в другом формате (например, «без двадцати пять»);
- проверка релевантности событию ДТП. Определяется, относится ли кандидат непосредственно к событию ДТП, а не к другим сущностям, описанным в тексте.

Запросы к LLM LLaMA строятся с учетом принципов промпт-инжиниринга для обеспечения точности и воспроизводимости ответов [17]. Структура запроса включает входные данные (исходный текст, описание кандидата), постановку задачи и ограничения на формат ответа: строго бинарный ответ («ДА» или «НЕТ») для минимизации неоднозначности его интерпретации.

В исследовании использовалась модель LLaMA-3-8B-Instruct в квантованном формате (4 bit, метод GGUF), развернутая локально. Для обеспечения воспроизводимости результатов и минимизации вариативности ответов значение параметра температуры было установлено равным 0. Риск галлюцинаций минимизируется за счет архитектуры метода: LLM выполняет функцию бинарного классификатора для кандидатов, уже найденных экспертными правилами. Такой подход ограничивает модель контекстом исходной публикации и конкретными упоминаниями.

Для формирования итогового результата устраняются конфликты, возникающие при наличии нескольких подтвержденных упоминаний даты и времени. Например, точное время («07:30») конкретизирует более общее упоминание («утром»). Если неоднозначность устранить не удастся или если после валидации не остается ни одного подтвержденного кандидата, предполагается, что дата и время события ДТП не могут быть извлечены.

3.1.2. Извлечение местоположения события ДТП. Для поиска упоминаний в тексте географических объектов используется географический справочник (газетир), который формируется для исследуемого региона на основе открытых данных проекта OpenStreetMap. В него включаются объекты двух категорий: дорожная сеть (улицы, проспекты, шоссе и т.д.) и объекты инфраструктуры (станции метро, торговые центры и т.д.). Для каждого объекта в газетире хранятся: наименование, тип и координаты (точки для объектов инфраструктуры и полилинии для элементов дорожной сети).

Помимо идентификации отдельных географических объектов, процедура учитывает пространственные отношения между ними, что позволяет более точно определять местоположение ДТП. К основным типам таких отношений относятся: пересечение улиц (например, «на пересечении Невского проспекта и Садовой улицы»), участок дороги (например, «на Лиговском проспекте между Кузнечным переулком и Разъездной улицей»), привязка к объекту-ориентир (например, «на Московском проспекте у станции метро Электросила»). Примеры таких правил, формально описанные с использованием нотации на основе Бэкуса-Наура (EBNF), в упрощенном виде представлены в листинге 1.

Извлечение местоположения события ДТП включает следующие последовательные этапы:

1. Предобработка текста. Данный этап аналогичен предобработке, выполняемой при извлечении даты и времени.

2. Идентификация географических объектов. Упоминания географических объектов выявляются путем нечеткого сопоставления N-грамм текста с наименованиями объектов в газетире. В качестве метрики сходства используется расстояние Левенштейна с пороговым значением коэффициента подобия $s > 0.85$. Для повышения точности также учитываются типы объектов (например, слово «улица»), указанные в тексте рядом с упоминанием названия объекта.

3. Валидация. Все найденные кандидаты проходят двухэтапную валидацию для исключения ложных совпадений. Во-первых, LLM LLaMA

подтверждает, что найденное упоминание действительно является географическим объектом. Во-вторых, проверяется релевантность этого объекта непосредственно местоположению события ДТП. Подход к построению запросов к LLM LLaMA соответствует описанному в подразделе об извлечении даты и времени.

4. Идентификация пространственных отношений. Для успешно валидированных географических объектов выполняется поиск упоминаний пространственных отношений между ними с использованием заранее сформированных экспертных правил. Если отношение найдено, его координаты вычисляются с помощью геометрии объектов из газетера. Координаты перекрестка определяются как точка пересечения полилиний соответствующих дорог; координаты участка дороги, как соответствующий сегмент полилинии основной дороги; координаты объекта относительно ориентира, как точка (проекция ориентира на дорогу).

В результате выполнения процедуры формируется набор координат (точек и полилиний), соответствующих всем идентифицированным и валидированным упоминаниям. Для получения итоговой точки применяется иерархический подход. Приоритет отдается точечным координатам, полученным из упоминаний отдельных объектов (например, станций метро), перекрестков или привязок к ориентирам. При наличии нескольких таких точек вычисляется их геометрический центр.

```
RELATION ::= INTERSECTION | ROAD_SEGMENT | LANDMARK_REF;

INTERSECTION ::=
    ("на" | "у"), ("пересечении" | "перекрестке"),
    ROADWAY, "и", ROADWAY;

ROAD_SEGMENT ::=
    ROADWAY, (
        ("между", (ROADWAY | PLACE), "и", (ROADWAY | PLACE))
        |
        ("от", (ROADWAY | PLACE), "до", (ROADWAY | PLACE))
    );

LANDMARK_REF ::=
    ("на" | "по"), ROADWAY, ("напротив" | "рядом с"),
    PLACE;
```

Листинг 1. Примеры экспертных правил для идентификации пространственных отношений

3.2. Функция схожести записей. Для количественной оценки соответствия двух записей $r_1, r_2 \in R$ одному событию ДТП на основе сравнения извлеченных значений их пространственно-временных атрибутов определяется функция схожести $sim : R \times R \rightarrow [0, 1]$.

Для построения функции вводятся два пороговых значения: τ_t (максимальное допустимое расхождение по времени) и τ_d (максимальное допустимое расхождение по расстоянию), при превышении которых записи не могут считаться эквивалентными. На их основе определяются две вспомогательные функции, оценивающие близость записей по времени ($sim_{datetime}$) и по расстоянию (sim_{space}):

$$sim_{datetime}(r_1, r_2) = \begin{cases} 1 - \frac{\delta_t(r_1, r_2)}{\tau_t}, & \delta_t(r_1, r_2) \leq \tau_t, \\ 0 & \end{cases} \quad (4)$$

где $\delta_t(r_1, r_2) = |a_{datetime}(r_1) - a_{datetime}(r_2)|$.

Функция схожести по расстоянию sim_{space} строится аналогично с использованием порогового значения τ_d и геодезического расстояния $\delta_d(r_1, r_2) = \|a_{loc_coords}(r_1) - a_{loc_coords}(r_2)\|$.

Итоговая функция схожести определяется как произведение $sim_{datetime}$ и sim_{space} . Построенная таким образом функция sim гарантирует, что схожесть записей будет высокой только в том случае, если они описывают события, которые произошли на небольшом расстоянии друг от друга и с незначительной временной разницей:

$$sim(r_1, r_2) = sim_{datetime}(r_1, r_2) \cdot sim_{space}(r_1, r_2). \quad (5)$$

Принимаются следующие формальные допущения об источниках шума в записях. Наблюдаемое значение атрибута $a(r)$ рассматривается как сумма истинного значения и ошибки (шума): $a(r) = \Theta_{e(r)} + \varepsilon(r)$. При этом возможны различные типы источников шума ε :

1. Для официальных записей ($R_{official}$) шум ε принимается пренебрежимо малым.

2. Для новостных публикаций (R_{news}) шум включает: неопределенность упоминаний даты и времени в тексте (использование неточных формулировок времени, например, «утром»),

погрешности методов извлечения атрибутов, а также возможные опечатки и фактические искажения в текстах публикаций.

Введение пороговых значений τ_t и τ_d для функции схожести позволяет компенсировать влияние данных источников шума при установлении эквивалентности записей.

Кусочно-линейный вид функций $sim_{datetime}$ и sim_{space} обусловлен необходимостью обеспечения устойчивости метода к зашумленным входным данным при извлечении атрибутов из неструктурированных текстов. В отличие от ступенчатых функций (жестких порогов), такой характер изменения схожести позволяет учитывать записи с умеренным уровнем $\varepsilon(r)$, не исключая их из процесса сопоставления, а пропорционально снижая значение схожести при увеличении расхождения. Это минимизирует риск ложного разделения кластеров при незначительных отклонениях в определении координат или времени события.

3.3. Алгоритм сопоставления записей. Для формирования искомого разбиения множества R на непересекающиеся кластеры, каждый из которых соответствует одному уникальному событию ДТП, предлагается алгоритм, осуществляющий распределение записей по кластерам на основе оценки близости значений их пространственно-временных атрибутов с использованием построенной функции схожести sim и введенных пороговых значений τ_t и τ_d .

Поскольку каждому официально зарегистрированному ДТП соответствует единственная запись в множестве $R_{official}$, а значения атрибутов $a_{datetime}$ и a_{loc_coords} для таких записей считаются точными, целесообразно рассматривать записи из $R_{official}$ как ядра формируемых кластеров. На первом этапе записи из R_{news} распределяются по данным кластерам на основе значений функции схожести sim с их ядрами, запись относится к тому кластеру, с ядром которого достигается максимальное положительное значение схожести. Если схожесть со всеми ядрами равна нулю, делается предположение, что запись описывает ДТП, сведения о котором не были официально опубликованы. На втором этапе такие нераспределенные записи подлежат кластеризации с использованием алгоритма ST-DBSCAN (модификации алгоритма DBSCAN, предназначенной для работы с пространственно-временными данными) [21]. Формальное описание алгоритма представлено в листинге 2.

Вход: $R = R_{\text{official}} \cup R_{\text{news}}$, пороговые значения τ_t, τ_d .
Выход: Разбиение $\mathcal{P} = \{R_1, \dots, R_l\}$ множества R .

Предварительный этап:

1. $C_{\text{init}} \leftarrow \{\{r\} \mid r \in R_{\text{official}}\}$ // Создание начальных кластеров
2. $R_{\text{unassigned}} \leftarrow \emptyset$

Этап 1: распределение записей из R_{news} по начальным кластерам

3. **for** each $r_{\text{news}} \in R_{\text{news}}$ **do**:
4. $r_{\text{official}}^* \leftarrow \arg \max_{r_{\text{official}} \in R_{\text{official}}} \text{sim}(r_{\text{news}}, r_{\text{official}})$
5. $\text{sim}_{\text{max}} \leftarrow \text{sim}(r_{\text{news}}, r_{\text{official}}^*)$
6. **if** $\text{sim}_{\text{max}} > 0$ **then**:
7. Найти кластер $R_i \in C_{\text{init}}$, такой что $r_{\text{official}}^* \in R_i$
8. $R_i \leftarrow R_i \cup \{r_{\text{news}}\}$
9. **else**:
10. $R_{\text{unassigned}} \leftarrow R_{\text{unassigned}} \cup \{r_{\text{news}}\}$
11. **end if**
12. **end for**

Этап 2: Кластеризация нераспределенных записей

13. // Параметры ST-DBSCAN:
14. $\text{eps}_{\text{space}} \leftarrow \tau_d$ // Максимальное расстояние между соседями
15. $\text{eps}_{\text{time}} \leftarrow \tau_t$ // Максимальная временная разница между соседями
16. $\text{min_samples} \leftarrow 2$ // Минимальное число соседей для формирования ядра кластера
17. $(C_{\text{clustered}}, R_{\text{noise}}) \leftarrow \text{ST-DBSCAN}(R_{\text{unassigned}}, \text{eps}_{\text{space}}, \text{eps}_{\text{time}}, \text{min_samples})$
18. $C_{\text{noise}} \leftarrow \{\{r\} \mid r \in R_{\text{noise}}\}$ // Каждая шумовая точка образует отдельный кластер
19. $\mathcal{P} \leftarrow C_{\text{init}} \cup C_{\text{clustered}} \cup C_{\text{noise}}$
20. **return** $\mathcal{P}$

Листинг 2. Алгоритм сопоставления записей

4. Экспериментальная оценка предлагаемого метода.

4.1. Методика экспериментальной оценки. Для экспериментальной оценки предлагаемого метода были собраны и размечены наборы данных о ДТП в Санкт-Петербурге за период с мая по июнь 2025 года. Эти наборы включают:

– Официальные карточки учета ДТП (R_{official}). Множество сформировано путем выгрузки данных с официального сайта Госавтоинспекции по всем ДТП, зарегистрированным в исследуемом регионе за указанный период.

– Новостные публикации (R_{news}). Множество сформировано на основе публикаций из шести источников: трех региональных

интернет-СМИ (в частности, «Фонтанка.ру») и трех тематических сообществ в социальной сети «ВКонтакте» (в частности, «ДТП и ЧП | Санкт-Петербург»). Данные были получены в автоматическом режиме с использованием методов веб-скрейпинга.

После завершения сбора записей была выполнена их экспертная разметка, включающая два этапа:

1. Для каждой новостной публикации из R_{news} вручную определены эталонные значения пространственно-временных атрибутов (a_{loc_coords} и $a_{datetime}$) на основе содержащихся в тексте указаний на местоположение и время события ДТП. В итоговые наборы данных включались только те записи, которые описывают одно конкретное ДТП в исследуемом регионе за указанный период и для которых удалось однозначно установить значения обоих атрибутов.

2. Выполнено ручное разбиение всего множества записей $R = R_{official} \cup R_{news}$ на кластеры. Разметка проведена на основе комплексного анализа содержания записей, включающего сопоставление сопроводительных фото- и видеоматериалов, выявление уникальных деталей в описании контекста ДТП, позволяющих с высокой степенью уверенности установить соответствие нескольких записей одному и тому же событию ДТП.

Разметка выполнялась двумя экспертами независимо. Согласованность аннотаторов (оценка через коэффициент каппа Козна κ) составила 0,85 для этапа извлечения атрибутов и 0,83 для этапа сопоставления записей. Выявленные расхождения разрешались совместно.

После завершения разметки собранные данные были разделены на два непересекающихся набора: валидационный набор (записи за май 2025 г.), предназначенный для отладки экспертных правил, настройки запросов к большой языковой модели и подбора пороговых значений τ_t и τ_d ; тестовый набор (записи за июнь 2025 г.), предназначенный для оценки предлагаемого метода. Характеристики наборов данных представлены в таблице 1.

Таблица 1. Характеристики наборов данных для экспериментальной оценки предложенного метода

Набор	Период записей	$ R_{official} $	$ R_{news} $	Кластеры
Валидационный	Май 2025	271	364	339
Тестовый	Июнь 2025	300	389	351

Для экспериментальной оценки выполнена программная реализация предложенного метода. На предварительном этапе единократно формируется локальный географический справочник

для региона Санкт-Петербург на основе данных платформы OpenStreetMap посредством запросов к интерфейсу Overpass API. Полученный справочник включает данные об элементах дорожной сети (улицы, проспекты, шоссе и др.) и объектах инфраструктуры (станции метро, торговые центры и др.). Для каждого объекта сохраняются его уникальный идентификатор, тип, наименование и географические координаты.

На этапе обработки записей, согласно предлагаемому методу, осуществляется анализ множества R_{news} для извлечения значений атрибутов $a_{datetime}$ и a_{loc_coords} из новостных публикаций, после чего выполняется построение разбиения множества R на кластеры. Реализация метода выполнена на языке программирования Python. Хранение данных, включая географический справочник (газетир) и записи о ДТП, организовано в реляционной базе данных PostgreSQL с использованием расширения PostGIS. Инференс большой языковой модели выполнен с использованием библиотеки llama-cpp-python ($temperature = 0$, фиксированный $seed$).

4.1.1. Подбор пороговых значений τ_t и τ_d . Работа алгоритма сопоставления записей (листинг 2) определяется двумя пороговыми значениями τ_t и τ_d . Для подбора значений был использован размеченный валидационный набор данных. Методика подбора основана на анализе распределения пространственно-временных расхождений между записями, которые, согласно эталонной разметке, описывают одно и то же ДТП.

Был проведен анализ чувствительности алгоритма к гиперпараметрам τ_t и τ_d . Исследование выполнялось путем изменения порогов в диапазоне от 85-го до 99-го перцентиля эмпирических распределений внутрикластерных расхождений на валидационном наборе. Установлено, что при значениях ниже 90-го перцентиля наблюдается резкое падение полноты, а выше 97-го перцентиля снижение точности вследствие ложного объединения событий с близкими значениями пространственно-временных атрибутов. При выбранном 95-й перцентиле малые отклонения параметров приводят к незначительным изменениям F1-меры.

Таким образом, в качестве итоговых пороговых значений были выбраны 95-е перцентили соответствующих эмпирических распределений внутрикластерных расхождений по времени и по расстоянию.

4.1.2. Метрики для оценки процедуры извлечения пространственно-временных атрибутов. Работа процедуры извлечения атрибутов a_{loc_coords} и $a_{datetime}$ из записей оценивается путем

сравнения извлеченных значений с эталонными значениями из тестового набора. Используются следующие метрики: полнота извлечения (доля записей, для которых процедура определила значение атрибута), средняя абсолютная ошибка (MAE) для оценки погрешности извлечения атрибута $a_{datetime}$ и средняя ошибка в расстоянии (MDE) для оценки погрешности извлечения атрибута a_{loc_coords} .

Для проверки гипотезы о том, что точность извлечения пространственно-временных атрибутов повышается за счет комбинации классических методов и большой языковой модели, способной интерпретировать сложные языковые конструкции, оценка процедуры извлечения выполняется в двух сценариях:

1. Использование исключительно экспертных правил и географического справочника (газетира).
2. Использование разработанной процедуры, объединяющей экспертные правила, географический справочник (газетир) и большую языковую модель LLaMA.

4.1.3. Метрики для оценки алгоритма сопоставления записей.

Работа алгоритма сопоставления записей (листинг 2) оценивается путем сравнения сформированного им разбиения записей на кластеры с эталонным разбиением записей тестового набора. Для этого используются метрики Точности (Precision), Полноты (Recall) и F1-меры (F1-score), которые вычисляются на основе анализа всех возможных пар записей в наборе. Оценка алгоритма сопоставления записей проводится в двух сценариях:

1. Алгоритм выполняется на тестовом наборе с использованием эталонных (размеченных вручную) значений пространственно-временных атрибутов для записей из R_{news} .
2. Алгоритм выполняется на тестовом наборе с использованием значений пространственно-временных атрибутов, извлеченных разработанной процедурой в сценарии с использованием LLM LLaMA.

4.2. Результаты экспериментальной оценки. На основе анализа эмпирических распределений внутрикластерных расхождений, построенных на валидационном наборе данных, определены пороговые значения τ_t и τ_d для алгоритма сопоставления записей (листинг 2). В качестве значений выбраны 95-е процентиля соответствующих распределений:

1. Максимальное допустимое расхождение по времени: $\tau_t = 4.2$ ч.
2. Максимальное допустимое расхождение по расстоянию: $\tau_d = 0.87$ км.

Оценка работы процедуры извлечения атрибутов $a_{datetime}$ и a_{loc_coords} из записей множества R_{news} для двух сценариев представлена в таблице 2.

Таблица 2. Оценка процедуры извлечения атрибутов из новостных публикаций на тестовом наборе данных

Атрибут	Полнота извлечения	Средняя ошибка
Сценарий 1 (без LLM)		
$a_{datetime}$	0.84	$MAE_{time} = 1.7$ ч.
a_{loc_coords}	0.81	$MDE_{space} = 0.52$ км
Сценарий 2 (с LLM LLaMA)		
$a_{datetime}$	0.8	$MAE_{time} = 1.1$ ч.
a_{loc_coords}	0.76	$MDE_{space} = 0.35$ км

Результаты оценки разбиения записей на кластеры, сформированного разработанным алгоритмом (листинг 2), представлены в таблице 3. Достигнутое значение F1-меры (0.84) признается достаточным для решения задачи сопоставления записей, так как оно соответствует уровню точности современных систем разрешения сущностей, работающих с текстовыми данными из неструктурированных источников [18].

Таблица 3. Оценка работы алгоритма сопоставления записей на тестовом наборе данных

Сценарий эксперимента	Precision	Recall	F1-score
С использованием эталонных значений атрибутов $a_{datetime}$ и a_{loc_coords}	0.96	0.91	0.93
С использованием извлеченных значений атрибутов $a_{datetime}$ и a_{loc_coords}	0.87	0.82	0.84

Для оценки устойчивости предложенного метода к ошибкам входных данных была проведена следующая проверка на тестовом наборе. В значения координат и времени случайно выбранных записей (20% от выборки) был внесен искусственный шум, имитирующий ошибки извлечения: смещение до 200 метров и 30 минут соответственно. Результаты показали, что точность сопоставления снизилась незначительно (F1-мера уменьшилась с 0,84 до 0,82), что подтверждает слабую чувствительность метода к незначительным погрешностям работы LLM и экспертных правил.

5. Обсуждение и заключение. В рамках исследования предложен метод решения задачи сопоставления записей о ДТП из официальных ресурсов и новостных публикаций в СМИ и социальных сетях на основе

комбинированной процедуры извлечения пространственно-временных атрибутов и алгоритма пространственно-временной кластеризации.

В отличие от детерминированных и вероятностных методов, предложенный метод позволяет эффективно обрабатывать неструктурированные тексты новостных публикаций без необходимости предварительного формирования полного набора структурированных атрибутов. В сравнении с методами на основе классического машинного обучения и глубокого обучения, предложенный метод не требует подготовки обширных размеченных наборов данных. Это существенно снижает трудозатраты на предварительную разметку данных о ДТП и обеспечивает возможность применения метода в различных регионах и смежных предметных областях без необходимости дообучения или сложной настройки моделей под специфику новых данных. В отличие от прямого использования больших языковых моделей для решения задачи сопоставления записей, комбинированный метод обеспечивает более высокую стабильность и интерпретируемость результатов. Поскольку роль большой языковой модели ограничена верификацией конкретных значений извлеченных атрибутов, а итоговое решение об эквивалентности записей принимается детерминированным алгоритмом кластеризации на основе пространственно-временных пороговых значений, минимизируется риск влияния галлюцинаций модели на корректность сопоставления записей.

Результаты экспериментальной оценки подтверждают обоснованность выдвинутых в исследовании гипотез. Высокие значения метрик алгоритма сопоставления в сценарии с использованием эталонных данных ($F1\text{-score} = 0,93$) подтверждают первую гипотезу: пространственно-временная близость описываемых в записях событий ДТП является подходящим критерием для установления эквивалентности записей о ДТП. Это обосновывает возможность использования пространственно-временных атрибутов в качестве ключевых признаков для разрешения сущностей в условиях отсутствия общих идентификаторов между гетерогенными источниками данных. Сравнительный анализ сценариев извлечения подтверждает вторую гипотезу: комбинирование классических методов с большой языковой моделью LLaMA повышает точность определения пространственно-временных атрибутов событий ДТП из новостных публикаций за счет семантической валидации упоминаний. Это подтверждает эффективность использования языковых моделей в качестве инструментов верификации для повышения качества извлечения данных из неструктурированных текстов на естественном

языке классическими методами, что позволяет обогащать доступные для анализа данные без необходимости подготовки обширных размеченных наборов для обучения моделей.

Предложенный метод обладает следующими ограничениями:

- Метод не применим к новостным публикациям, содержащим описание нескольких событий ДТП, а также к публикациям, в которых отсутствуют достаточные данные для определения местоположения и времени ДТП;

- Эффективность сопоставления записей существенно зависит от точности работы процедуры извлечения пространственно-временных атрибутов из текстов;

- Эффективность сопоставления записей ограничена полнотой и актуальностью данных OpenStreetMap для исследуемого региона;

- В условиях аномально высокой плотности событий (например, несколько ДТП на одном перекрестке в пределах одного часа) использование исключительно пространственно-временных атрибутов может приводить к ошибочному объединению записей в один кластер;

- Использование большой языковой модели в процедуре извлечения данных накладывает ограничения на скорость обработки потока записей в режиме реального времени из-за значительных требований к вычислительным ресурсам;

- Возможно возникновение галлюцинаций в большой языковой модели. Однако использование локального географического справочника и экспертных правил на этапе формирования кандидатов позволяет минимизировать данный риск, ограничивая область работы LLM верификацией только реально существующих объектов исследуемого региона.

Дальнейшие исследования могут вестись в следующих направлениях:

- Разработка алгоритма декомпозиции текстов, описывающих несколько событий ДТП, для их последующего отдельного сопоставления с использованием предложенного метода;

- Автоматизация процесса обогащения официальных карточек учета ДТП дополнительными данными из кластеризованных новостных публикаций, включая фотоматериалы и описание причин ДТП, для формирования единого расширенного контекста событий;

- Проведение анализа воспроизводимости результатов и экспериментальная оценка адаптивности метода на данных из других регионов и смежных предметных областей. В рамках данного направления предполагается уточнение теоретических границ применимости метода,

в частности, определение критических значений плотности событий в пространстве и времени, при которых сохраняется высокая точность сопоставления записей;

– Создание на основе предложенного метода полнофункциональной аналитической системы, обеспечивающей автоматизированное извлечение записей из первичных источников, определение значений пространственно-временных атрибутов, построение разбиения множества записей на кластеры и визуализацию результатов интеграции данных.

Литература

1. Официальный сайт Госавтоинспекции МВД России. URL: <https://stat.gibdd.ru/> (дата обращения: 01.12.2025).
2. Rabbani M.B.A., Musarat M.A., Alaloul W., et al. Road accident data collection systems in developing and developed countries: a review // *International Journal of Integrated Engineering*. 2022. vol. 14. no. 1. pp. 336–352. DOI: 10.30880/ijie.2022.14.01.031.
3. Chand A., Jayesh S., Bhasi A.B. Road traffic accidents: An overview of data sources, analysis techniques and contributing factors // *Materials Today: Proceedings*. 2021. vol. 47. pp. 5135–5141. DOI: 10.1016/j.matpr.2021.05.415.
4. Dos Santos S.R., Jr Davis C.A., Smarzarzo R. Integration of data sources on traffic accidents // *Proceedings of the XVII GEOINFO*. 2016. pp. 93–104.
5. Ali F., Ali A., Imran M., et al. Traffic accident detection and condition analysis based on social networking data // *Accident Analysis Prevention*. 2021. vol. 151. 105973 p. DOI: 10.1016/j.aap.2021.105973.
6. Chang H., Li L., Huang J., et al. Tracking traffic congestion and accidents using social media data: A case study of Shanghai // *Accident Analysis Prevention*. 2022. vol. 169. 106618 p. DOI: 10.1016/j.aap.2022.106618.
7. Митряев И.С., Новиков А.Н., Кравченко А.А., Еремин С.В. Методы повышения точности извлечения информации из социальных сетей для интеллектуальных транспортных систем // *Мир транспорта и технологических машин*. 2024. №4-2(87). С. 114–121.
8. Girin A., Teslya N., Shilov N. Overview of Publicly-Available Data Sources on Road Traffic Accidents in Russia // *Proceedings of the 10th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS)*. 2024. pp. 480–487. DOI: 10.5220/0012737100003702.
9. Fellegi I.P., Sunter A.B. A Theory for Record Linkage // *Journal of the American Statistical Association*. 1969. vol. 64. no. 328. pp. 1183–1210. DOI: 10.1080/01621459.1969.10501049.
10. Binette O., Steorts R.C. (Almost) all of entity resolution // *Science Advances*. 2022. vol. 8. no. 12. eabi8021 p. DOI: 10.1126/sciadv.abi8021.
11. Elmagarmid A.K., Ipeirotis P.G., Verykios V.S. Duplicate Record Detection: A Survey // *IEEE Transactions on Knowledge and Data Engineering*. 2007. vol. 19. no. 1. pp. 1–16. DOI: 10.1109/TKDE.2007.250581.
12. Mudgal S., Li H., Rekatsinas T., et al. Deep Learning for Entity Matching: A Design Space Exploration // *Proceedings of the International Conference on Management of Data*. 2018. pp. 19–34.
13. Li Y., Li J., Suhara Y., et al. Effective entity matching with transformers // *The VLDB Journal*. 2023. vol. 32. pp. 1215–1235. DOI: 10.1007/s00778-023-00779-z.

14. Li Q., Li J., Sheng J., et al. A Survey on Deep Learning Event Extraction: Approaches and Applications // *IEEE Transactions on Neural Networks and Learning Systems*. 2024. vol. 35. no. 5. pp. 6301–6321. DOI: 10.1109/TNNLS.2022.3213168.
15. Paganelli M., Tiano D., Del Buono F., et al. How Transformers are Revolutionizing Entity Matching // *Proceedings of the 32nd Symposium on Advanced Database Systems*. 2024. vol. 3741. pp. 662–670.
16. Peeters R., Steiner A., Bizer C. Entity Matching using Large Language Models // *arXiv preprint arXiv:2310.11244*. 2023.
17. Marvin G., Hellen N., Jjing D., et al. Prompt Engineering in Large Language Models // *Algorithms for Intelligent Systems*. Springer. 2023. pp. 387–402. DOI: 10.1007/978-981-99-7962-2_30.
18. Peeters R., Bizer C. Using ChatGPT for Entity Matching // *Communications in Computer and Information Science*. Springer. 2023. vol. 1850. pp. 221–230. DOI: 10.1007/978-3-031-42941-5_20.
19. Li H., Li S., Hao F., et al. BoostER: Leveraging Large Language Models for Enhancing Entity Resolution // *Companion Proceedings of the ACM Web Conference*. 2024. pp. 1043–1046. DOI: 10.1145/3589335.3651245.
20. Touvron H., Lavril T., Izacard G., et al. LLaMA: Open and Efficient Foundation Language Models // *arXiv preprint arXiv:2302.13971*. 2023.
21. Birant D., Kut A. ST-DBSCAN: An algorithm for clustering spatial–temporal data // *Data Knowledge Engineering*. 2007. vol. 60. no. 1. pp. 208–221. DOI: 10.1016/j.datak.2006.01.013.
22. Strotgen J., Gertz M. Heildeltime: High quality rule-based extraction and normalization of temporal expressions // *Proceedings of the 5th International Workshop on Semantic Evaluation*. 2010. pp. 321–324.
23. Giridhar P., Abdelzaher T., George J., et al. On quality of event localization from social network feeds // *Proceedings of the IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. 2015. pp. 75–80. DOI: 10.1109/PERCOMW.2015.7133997.
24. Milusheva S., Marty R., Bedoya G., et al. Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning // *PLOS One*. 2021. vol. 16. no. 2. e0244317 p. DOI: 10.1371/journal.pone.0244317.
25. Suat-Rojas N., Gutierrez-Osorio C., Pedraza C. Extraction and Analysis of Social Networks Data to Detect Traffic Accidents // *Information*. 2022. vol. 13. no. 1. 26 p. DOI: 10.3390/info13010026.
26. Hu X., Zhou Z., Li H., et al. Location Reference Recognition from Texts: A Survey and Comparison // *ACM Computing Surveys*. 2023. vol. 56. no. 5. pp. 1–37. DOI: 10.1145/3625819.
27. Xiang W., Wang B. A Survey of Event Extraction From Text // *IEEE Access*. 2019. vol. 7. pp. 173111–173137. DOI: 10.1109/ACCESS.2019.2956831.
28. Hu Y., Mai G., Cundy C., et al. Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages // *International Journal of Geographical Information Science*. 2023. vol. 37. no. 11. pp. 2289–2318. DOI: 10.1080/13658816.2023.2266495.

Гирин Алексей Романович — аспирант, Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО» (ИТМО). Область научных интересов: интеграция данных, обработка текста на естественном языке, языковые модели. Число научных публикаций — 8.

agirin@itmo.ru; Кронверкский пр., 49А, 197101, Санкт-Петербург, Россия; р.т.: +7(812)480-0370.

Тесля Николай Николаевич — канд. техн. наук, старший научный сотрудник, лаборатория интегрированных систем автоматизации, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН). Область научных интересов: управление знаниями, онтологии, обработка текста на естественном языке, языковые модели. Число научных публикаций — 100. teslya@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)508-3311.

Поддержка исследований. Работа выполнена в рамках бюджетной темы FFZF-2025-0003.

A. GIRIN, N. TESLYA
**RECORD LINKAGE METHOD FOR ROAD TRAFFIC ACCIDENTS
USING OFFICIAL DATA SOURCES AND NEWS PUBLICATIONS
BASED ON SPATIO-TEMPORAL CLUSTERING**

Girin A., Teslya N. Record Linkage Method for Road Traffic Accidents Using Official Data Sources and News Publications Based on Spatio-Temporal Clustering.

Abstract. Effective decision-making for reducing the road accident rate in motor transport is based on the results of analyzing traffic accident hotspots. The completeness and accuracy of accident data used for analysis can be enhanced through the integration of multiple sources. A crucial step in integration is the matching of records from different sources that describe the same traffic accident events. Solving the matching problem is complicated by heterogeneity: accident records may differ in structure and format. This study proposes a method for matching accident records from two heterogeneous sources: official resources with structured traffic accident report cards and news publications in the form of unstructured natural language texts. The proposed method is based on the hypothesis of spatial and temporal proximity: records from different sources are considered to correspond to the same accident if the events described occurred within a short distance and with a negligible time difference. The method comprises two sequential stages. At the first stage, coordinates and date/time of the described event are extracted from news publications using expert rules, a geographic reference database, and the LLaMA large language model. At the second stage, records from official resources and news publications are matched using spatiotemporal clustering with the ST-DBSCAN algorithm. The method enables the enrichment of the contextual description of officially registered traffic accidents and enhances the data available for analysis by incorporating accidents not recorded in official sources. Experimental evaluation of the method was conducted on real-world traffic accident data from Saint Petersburg. The results demonstrated that combining classical methods with a large language model improves the accuracy of extracting spatiotemporal attributes of traffic accidents from news publications, while spatiotemporal clustering ensures high accuracy in matching accident records (F1-score = 0.84).

Keywords: data integration, record linkage, traffic accidents, gazetteer, expert rules, ST-DBSCAN, large language models, LLaMA.

References

1. Oficial'nyj sajt Gosavtoinspektzii MVD Rossii [Official web site of the State Traffic Inspectorate of the Ministry of Internal Affairs of Russia]. Available at: <https://stat.gibdd.ru/> (accessed 01.12.2025). (In Russ.).
2. Rabbani M.B.A., Musarat M.A., Alaloul W., et al. Road accident data collection systems in developing and developed countries: a review. *International Journal of Integrated Engineering*. 2022. vol. 14. no. 1. pp. 336–352. DOI: 10.30880/ijie.2022.14.01.031.
3. Chand A., Jayesh S., Bhasi A.B. Road traffic accidents: An overview of data sources, analysis techniques and contributing factors. *Materials Today: Proceedings*. 2021. vol. 47. pp. 5135–5141. DOI: 10.1016/j.matpr.2021.05.415.
4. Dos Santos S.R., Jr Davis C.A., Smarzaro R. Integration of data sources on traffic accidents. *Proceedings of the XVII GEOINFO*. 2016. pp. 93–104.

5. Ali F., Ali A., Imran M., et al. Traffic accident detection and condition analysis based on social networking data. *Accident Analysis Prevention*. 2021. vol. 151. 105973 p. DOI: 10.1016/j.aap.2021.105973.
6. Chang H., Li L., Huang J., et al. Tracking traffic congestion and accidents using social media data: A case study of Shanghai. *Accident Analysis Prevention*. 2022. vol. 169. 106618 p. DOI: 10.1016/j.aap.2022.106618.
7. Mitryaev I.S., Novikov A.N., Kravchenko A.A., Eremin S.V. [Methods for improving the accuracy of information extraction from social networks for intelligent transport systems]. *Mir transporta i tekhnologicheskikh mashin – The world of transport and technological machines*. 2024. no. 4-2(87). pp. 114–121. (In Russ.).
8. Girin A., Teslya N., Shilov N. Overview of Publicly-Available Data Sources on Road Traffic Accidents in Russia. *Proceedings of the 10th International Conference on Vehicle Technology and Intelligent Transport Systems (VEHITS)*. 2024. pp. 480–487. DOI: 10.5220/0012737100003702.
9. Fellegi I.P., Sunter A.B. A Theory for Record Linkage. *Journal of the American Statistical Association*. 1969. vol. 64. no. 328. pp. 1183–1210. DOI: 10.1080/01621459.1969.10501049.
10. Binette O., Steorts R.C. (Almost) all of entity resolution. *Science Advances*. 2022. vol. 8. no. 12. eabi8021 p. DOI: 10.1126/sciadv.abi8021.
11. Elmagarmid A.K., Ipeirotis P.G., Verykios V.S. Duplicate Record Detection: A Survey. *IEEE Transactions on Knowledge and Data Engineering*. 2007. vol. 19. no. 1. pp. 1–16. DOI: 10.1109/TKDE.2007.250581.
12. Mudgal S., Li H., Rekatsinas T., et al. Deep Learning for Entity Matching: A Design Space Exploration. *Proceedings of the International Conference on Management of Data*. 2018. pp. 19–34.
13. Li Y., Li J., Suhara Y., et al. Effective entity matching with transformers. *The VLDB Journal*. 2023. vol. 32. pp. 1215–1235. DOI: 10.1007/s00778-023-00779-z.
14. Li Q., Li J., Sheng J., et al. A Survey on Deep Learning Event Extraction: Approaches and Applications. *IEEE Transactions on Neural Networks and Learning Systems*. 2024. vol. 35. no. 5. pp. 6301–6321. DOI: 10.1109/TNNLS.2022.3213168.
15. Paganelli M., Tiano D., Del Buono F., et al. How Transformers are Revolutionizing Entity Matching. *Proceedings of the 32nd Symposium on Advanced Database Systems*. 2024. vol. 3741. pp. 662–670.
16. Peeters R., Steiner A., Bizer C. Entity Matching using Large Language Models. *arXiv preprint arXiv:2310.11244*. 2023.
17. Marvin G., Hellen N., Jjing D., et al. Prompt Engineering in Large Language Models. *Algorithms for Intelligent Systems*. Springer. 2023. pp. 387–402. DOI: 10.1007/978-981-99-7962-2_30.
18. Peeters R., Bizer C. Using ChatGPT for Entity Matching. *Communications in Computer and Information Science*. Springer. 2023. vol. 1850. pp. 221–230. DOI: 10.1007/978-3-031-42941-5_20.
19. Li H., Li S., Hao F., et al. BoostER: Leveraging Large Language Models for Enhancing Entity Resolution. *Companion Proceedings of the ACM Web Conference*. 2024. pp. 1043–1046. DOI: 10.1145/3589335.3651245.
20. Touvron H., Lavril T., Izacard G., et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*. 2023.
21. Birant D., Kut A. ST-DBSCAN: An algorithm for clustering spatial–temporal data. *Data Knowledge Engineering*. 2007. vol. 60. no. 1. pp. 208–221. DOI: 10.1016/j.datak.2006.01.013.

22. Strotgen J., Gertz M. Heildeltime: High quality rule-based extraction and normalization of temporal expressions. *Proceedings of the 5th International Workshop on Semantic Evaluation*. 2010. pp. 321–324.
23. Giridhar P., Abdelzaher T., George J., et al. On quality of event localization from social network feeds. *Proceedings of the IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*. 2015. pp. 75–80. DOI: 10.1109/PERCOMW.2015.7133997.
24. Milusheva S., Marty R., Bedoya G., et al. Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning. *PLOS One*. 2021. vol. 16. no. 2. e0244317 p. DOI: 10.1371/journal.pone.0244317.
25. Suat-Rojas N., Gutierrez-Osorio C., Pedraza C. Extraction and Analysis of Social Networks Data to Detect Traffic Accidents. *Information*. 2022. vol. 13. no. 1. 26 p. DOI: 10.3390/info13010026.
26. Hu X., Zhou Z., Li H., et al. Location Reference Recognition from Texts: A Survey and Comparison. *ACM Computing Surveys*. 2023. vol. 56. no. 5. pp. 1–37. DOI: 10.1145/3625819.
27. Xiang W., Wang B. A Survey of Event Extraction From Text. *IEEE Access*. 2019. vol. 7. pp. 173111–173137. DOI: 10.1109/ACCESS.2019.2956831.
28. Hu Y., Mai G., Cundy C., et al. Geo-knowledge-guided GPT models improve the extraction of location descriptions from disaster-related social media messages. *International Journal of Geographical Information Science*. 2023. vol. 37. no. 11. pp. 2289–2318. DOI: 10.1080/13658816.2023.2266495.

Girin Alexey — Graduate Student, ITMO University. Research interests: data integration, natural language text processing, language models. The number of publications — 8. agirin@itmo.ru; 49A, Kronverkskiy Ave., 197101, St. Petersburg, Russia; office phone: +7(812)480-0370.

Teslya Nikolay — Ph.D., Senior Researcher, Laboratory of Integrated Automation Systems, St. Petersburg Federal Research Center of the Russian Academy of Sciences. Research interests: knowledge management, ontologies, natural language text processing, language models. The number of publications — 100. teslya@ias.spb.su; 39, 14-th Line V.O., 199178, St. Petersburg, Russia; office phone: +7(812)508-3311.

Acknowledgements. The work was supported out within the framework of the budget theme FFZF-2025-0003.