УДК 004.912

ПОДХОД К ЛОКАЛИЗАЦИИ МЕСТ ДТП ИЗ ТЕКСТОВ НОВОСТЕЙ С ИСПОЛЬЗОВАНИЕМ ГАЗЕТИРА И БОЛЬШОЙ ЯЗЫКОВОЙ МОДЕЛИ LLAMA

Гирин А.Р. (ИТМО)
Научный руководитель – кандидат технических наук, доцент Тесля Н.Н.
(ИТМО)

Введение. Средства массовой информации являются значимым источником данных о дорожно-транспортных происшествиях (ДТП). Сведения о ДТП, полученные из новостных текстов, позволяют экспертам в области управления дорожным трафиком оперативно реагировать на ДТП и предупреждать водителей о заторах. Однако новостные тексты, как правило, представлены в неструктурированном виде. Это затрудняет их использование в системах мониторинга дорожного трафика без предварительной обработки. Важным этапом такой обработки является локализация места ДТП, описанного в тексте [1].

Задача локализации мест ДТП из новостных текстов является частным случаем задачи локализации мест событий, описанных в неструктурированных текстах. Для решения этой задачи разработано несколько классов методов: но основе правил; на основе газетира; на основе машинного обучения [2]. В настоящей работе проведен обзор этих методов. Предложен комбинированный подход с использованием этих методов для решения исходной задачи локализации места ДТП из новостных текстов. Подход апробирован на новостных текстах по теме ДТП в г. Санкт-Петербурге.

Методы на основе правил предполагают ручное создание набора правил для поиска упоминаний мест в тексте. Эти правила часто опираются на части речи, ключевые слова, указывающие на типы географических объектов, и используют регулярные выражения или контекстно-свободные грамматики для определения, является ли N-грамма из текста упоминанием места. Основные ограничения таких методов: высокий уровень ложноположительных срабатываний и сложность создания всеохватывающих правил [3].

Методы на основе газетира используют словари географических названий, связанных с координатами и типами объектов. N-граммы из текста сравниваются с записями газетира, а совпадения принимаются за упоминание места. Этот подход имеет ряд проблем: возможный пропуск упоминаний из-за вариативности названий или неполноты газетира; неоднозначность, вызванная совпадением названий из газетира с негеографическими сущностями или дублированием названий для разных объектов [4]. Эффективность метода повышается при составлении газетира для небольшой территории (например, города) [2].

Поиск упоминаний мест в тексте является частным случаем задачи распознавания именованных сущностей (NER). Для идентификации мелкомасштабных объектов, таких как улицы, требуется дообучение существующих моделей NER на специализированных данных, что делает такой подход ресурсозатратным. Отдельно стоит отметить большие языковые модели (LLM), которые способны учитывать контекст и семантику текста, что делает их перспективными для работы с неформальными неструктурированными текстами [2].

Таким образом, методы, основанные только на правилах, малоэффективны и обычно комбинируются с другими методами. Методы на основе газетира требуют значительного числа эвристических правил для анализа контекста и уменьшения неоднозначности, а также для проверки, относятся ли найденные места к конкретному событию. Использование моделей NER эффективно, но требует дообучения на размеченных данных, что увеличивает сложность и ресурсоемкость подхода. Использование LLM для решения задачи локализации места события в тексте является перспективным направлением и требует развития [2].

Основная часть. Существенной проблемой при использовании газетира является необходимость ввода сложных правил для уточнения, является ли найденное совпадение упоминанием места (следствие неоднозначности наименований). В настоящей работе предлагается валидировать найденные совпадения через запросы к LLM, способной

эффективно анализировать контекст возможных упоминаний мест в тексте [2]. Также предлагается выявлять в тексте пространственные отношения между объектами (например, пересечения) для уточнения места события. Для этого вводятся дополнительные правила.

Для решения задачи локализации мест ДТП из новостных текстов предлагается комбинированный подход, состоящий из нескольких этапов.

Этап 1. Подготовка газетира для исследуемой области с использованием данных OpenStreetMap. Объекты классифицируются на три категории: территории (районы); проезды (улицы, проспекты и др.); места (дома, станции метро, ТРЦ и др.). Для каждого объекта фиксируются название, координаты, тип.

Этап 2. Предобработка текста: очистка от ссылок, токенизация и лемматизация.

Этап 3. С использованием N-грамм осуществляется поиск упоминаний в тексте объектов с названиями из газетира через нечеткое сравнение строк (расстояние Джаро-Винклера, порог ≥ 0.85). Дополнительно вводится правило проверки типа объекта через фиксированные суффиксы/префиксы («ул.», «улица.»). Каждому найденному упоминанию присваивается вес, рассчитываемый как комбинация метрик схожести и выполнения правила.

Этап 4. Для снижения ложноположительных поисков применяется LLM Llama-3.2-3B-Instruct. Модель подтверждает наличие упоминания найденных объектов в тексте с учетом контекста и проверяет связь места объекта с местом ДТП. Ответы модели («да/нет») корректируют веса объектов. Объекты с весом <0.7 исключаются.

Этап 5. Через правила на основе регулярных выражений в тексте выявляются зависимости (отношения) между найденными объектами, а также рассчитываются координаты таких отношений: наличие в тексте шаблонов вида «перекресток улиц X и Y» или «поворот с улицы X на Y» (координатой отношения является точка пересечения улиц) и других. Вес отношений рассчитывается на основе весов входящих в него объектов.

Этап 6. Все найденные кандидаты (объекты и отношения) ранжируются по весам. На выход подается список координат с указанием веса в качестве уровня достоверности.

Для апробации подхода были использованы 40 размеченных новостных текстов о ДТП в г. Санкт-Петербурге, а также составлен газетир для объектов города. Точность определения мест (без Llama): 65%. Точность определения мест (с Llama): 80%.

Выводы. Предложен эффективный комбинированный подход для локализации мест ДТП из новостных текстов. Использование локального газетира позволило минимизировать неоднозначности при поиске упоминаний мест в тексте, а интеграция с LLM Llama повысила точность проверки найденных совпадений. Подход применим при сборе данных для анализа ДТП, но также может использоваться и в других областях для решения аналогичных задач.

Список использованных источников:

1. Ali F. et al. Traffic accident detection and condition analysis based on social networking data //Accident Analysis & Prevention. – 2021. – Т. 151. – С. 105973.
2. Hu X. et al. Location reference recognition from texts: A survey and comparison //ACM Computing Surveys. – 2023. – Т. 56. – №. 5. – С. 1-37.
3. Giridhar P. et al. On quality of event localization from social network feeds //2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops). – IEEE, 2015. – С. 75-80.
4. Milusheva S. et al. Applying machine learning and geolocation techniques to social media data (Twitter) to develop a resource for urban planning //PloS one. – 2021. – Т. 16. – №. 2. – С. e0244317.