УДК: 528.926:004 **DOI: 10.35595/2414-9179-2022-1-28-659-669**

А.А. Колесников[1], Е.А. Плитченко[2], М.К. Кропачева[3]

АВТОМАТИЗАЦИЯ ПОДГОТОВКИ КАРТОГРАФИЧЕСКИХ ДАННЫХ С ПОМОЩЬЮ СИСТЕМ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

АННОТАЦИЯ

Современный уровень развития информационных технологий позволяет автоматизировать обработку тех видов данных, с которыми ранее мог работать только специалист. В качестве одного из таких примеров можно привести технологии обработки естественного языка, реализующие функции анализа тональности, машинного перевода, вопросно-ответных систем. Для процессов создания картографических и геоинформационных произведений наибольший интерес представляют методики извлечения именованных сущностей, которые позволяют извлекать географические названия из неструктурированного текста, и связывания именованных сущностей, дающие возможность создания логических связей между извлеченными наименованиями пространственных объектов. Их обработка посредством локальной или сетевой базы данных сервиса для геокодирования позволит автоматизировать создание слоев карты в геоинформационной системе на основе текстовых сообщений. В статье описываются наиболее популярные подходы и их программные реализации для решения задачи извлечения именованных сущностей на примере текстов биографий и произведений сибирских писателей. Выполняется анализ методик, основанных на правилах, моделях максимальной энтропии и сверточных нейронных сетях. Для оценки качества результатов извлечения из текста географических названий и объектов, помимо стандартного варианта F1-score, авторами предлагается дополнительный вариант способа оценки, учитывающий большее число критериев и также базирующийся на матрице ошибок. Приведено описание форматов разметки текстовых блоков для улучшения качества распознавания и расширения возможных вариантов географических наименований именованных сущностей на основе дообучения модели нейронной сети.

КЛЮЧЕВЫЕ СЛОВА: географическое название, автоматизация, извлечение именованных сущностей, обработка естественного языка, нейронные сети, сибирские писатели

[1] Сибирский государственный университет геосистем и технологий, ул. Плахотного, д. 10, 630108, Новосибирск, Россия *e-mail:* **alexeykw@mail.ru**

[2] Фонд поддержки литературного творчества «Сибирский писатель», ул. Грибоедова, д. 2, офис 11, 630083, Новосибирск, Россия *e-mail:* **str2007@list.ru**

[3] Сибирский государственный университет геосистем и технологий, ул. Плахотного, д. 10, 630108, Новосибирск, Россия e-mail: **kropacheva.m.k@gmail.com**

Alexey A. Kolesnikov[1], **Egor A. Plitchenko**[2], **Maria K. Kropacheva**[3]

AUTOMATION OF DATA PREPARATION FOR MAPPING USING NATURAL LANGUAGE PROCESSING SYSTEMS

ABSTRACT

The current level of development of information technology makes it possible to automate the processing of those types of data that only a specialist could previously work with. One such example is natural language processing technologies that implement the functions of sentiment analysis, machine translation, and question-answer systems. For the processes of creating cartographic and geoinformation works, the methods of extracting named entities are of the greatest interest, which allows extracting geographical names from unstructured text and linking named entities, which make it possible to create logical links between the extracted names of spatial objects. Their processing, through a local or network database of the service for geocoding, will automate the creation of map layers in a geographic information system based on text messages. The article describes the most popular approaches and their software implementations for solving the problem of extracting named entities in the example of texts of biographies and works of Siberian writers. Rule-based methodologies, maximum entropy models, and convolutional neural networks are analyzed. To assess the quality of the results of extracting geographical names and objects from the text, in addition to the standard F1-score, the authors propose an additional variant of the evaluation method that takes into account a larger number of criteria and is also based on an error matrix. The description of text block markup formats is given to improve the quality of recognition and expand the possible options for geographical names of named entities based on additional training of the neural network model.

KEYWORDS: geographical name, automation, named entity extraction, natural language processing, neural networks, Siberian writers

ВВЕДЕНИЕ

Писатели Сибири внесли значительный вклад в культуру России. Всеволод Иванов, Николай Гарин-Михайловский, Всеволод Гаршин, Афанасий Коптелов, Михаил Михеев, Владимир Сапожников, Александр Денисенко, Александр Плитченко и многие другие, вплоть до современности, в своих произведениях описывают Сибирь, раскрывая тем самым многообразие культуры России, показывая посредством местного колорита общие для страны социальные и культурные явления, формируя героями своих произведений пример для воспитания искренне преданных родине, честных и трудолюбивых граждан. В ходе опроса в социальной сети ВКонтакте был сделан вывод, что у современной молодежи низкий интерес к отечественной литературе, истории и родному языку. Именно здесь видится важная роль отечественной культуры и литературы, в частности. Современный культурный процесс развивается вместе с технологиями, активно входят в жизнь цифровые формы, совершившие переворот в визуальном искусстве, как когда-то кинематограф. Отчетливо видна важная роль цифровизации в применении современных технологий в мире культуры. Для того чтобы донести до молодежи самые важные, самые сокровенные мысли, необходимо говорить с ней языком тех образов и

[1] Siberian State University of Geosystems and Technologies, Plakhotnogo str., 10, 630108, Novosibirsk, Russia, e-mail*:* **alexeykw@mail.ru**

[2] Foundation for Support of Literary Creativity “Siberian Writer”, Griboyedova str., 2-11, 630083, Novosibirsk, Russia, *e-mail:* **str2007@list.ru**

[3] Siberian State University of Geosystems and Technologies, Plakhotnogo str., 10, 630108, Novosibirsk, Russia, e-mail: **kropacheva.m.k@gmail.com**

теми средствами, которые близки и востребованы, использовать доступные и понятные способы. Например, такими формами являются интерактивные картографические произведения, используя которые можно реализовать следующую задачу – воспитать, показать наглядно, рассказать о сибирских писателях и их книгах, заинтересовать и привлечь молодежь посредством понятного и интересного контента к изучению родной культуры и истории. Актуальность русской сибирской литературы и литераторов, фактов их биографий сегодня высока. Использование этих материалов в качестве основы для создания картографических произведений мы видим актуальным как при создании интуитивно понятного, близкого пользователям контента, так и для побуждения молодежи к саморазвитию посредством изучения русской сибирской литературы. Связующим звеном между текстовыми данными и их пространственным представлением могут стать технологии обработки и анализа естественного языка.

Технологии обработки естественного языка (англ. Natural Language Processing, NLP) сейчас все более широко внедряются в самые разные области деятельности и позволяют решать различные задачи, включая машинный перевод, распознавание речи, чат-боты, вопросно-ответные системы, автореферирование, генерацию текстов, анализ тональности, системы рекомендаций и поиска информации.

Технологии NLP базируются на междисциплинарном подходе, который опирается на лингвистику, информатику, статистику и другие смежные дисциплины, включая географию и геоинформатику. В последние годы в этой области значительное внимание уделялось извлечению сущностей из больших объемов данных на естественном языке, таких как текст и аудиоролики. Сегодня это направление исследований сместилось к семантическому «пониманию» и связыванию понятий между собой с целью ответа на вопросы и контекстного информирования при принятии решений [Карпачевский*, Филиппова*, 2018; *Cooper* et al., 2016].

Изначально к задачам обработки естественного языка относилось извлечение структур из неструктурированного текста. Ранние подходы к интерпретации текста и речи основывались на определяемых пользователем наборах правил. Затем, в 1980-х и 1990-х гг. были разработаны методы статистического вывода для выявления и применения этих правил к естественному языку. Развитие технологий искусственного интеллекта привело к широкому использованию методов машинного обучения и нейросетевых технологий, включая глубокое обучение. Они ориентируются не на использование правил, а направлены на «понимание» естественного языка с помощью статистических методов и представления текста в виде семантических сетей, которые могут идентифицировать лингвистические свойства слов, предложений или документов [*Bodenhamer* et al., 2015; *Cura* et al., 2018].

Хотя NLP-технологии могут применяться в большинстве сфер деятельности человека, значительная часть содержания человеческого языка находится в прямой или косвенной связи с географическим пространством и временем. Кроме этого, естественный язык различается в зависимости от территории, а это означает, что ГИС-специалисты могут использовать свой опыт для обработки, идентификации и контекстуализации языковых шаблонов. Учитывая это, для более корректного извлечения географических названий из текстов, собранных из разных источников (особенно в случае использования разных языков), требуется не только обучить математическую модель, но и дополнить обучающую выборку и базу данных сервиса геокодирования дополнительными топонимами [*Akbik* et al., 2018; *Berant* et al., 2013]. В области геоинформационных технологий NLP используется для лучшего понимания и установления взаимосвязей для широкого спектра географических объектов, посредством идентификации мест, событий и явлений, а также извлечения лингвистических характеристик, связанных с этими объектами. Также методы NLP могут поспособствовать более глубокому пониманию географических процессов и явлений, которое может быть недоступно с помощью традиционного пространственно-временного анализа. Знание пространственных отношений, региональных иерархий, географических законов и теорий в сочетании со многими ведущими

методами и технологиями обработки естественного языка приводит к появлению передовых инструментов и ГИС-приложений, многие из которых активно развиваются и используются сегодня в науке и производстве [*Anh* et al., 2017; *Ding* et al., 2018; *Mozharova, Loukachevitch*, 2016].

Применение существующего инструментария NLP совместно с ГИС-технологиями позволяет извлекать географическую информацию из отчетов, новостных лент, книг и статей, постов социальных сетей, и формировать на основе этого отдельные слои геоинформационной модели и тематические карты. Кроме этого, используя принципы связывания именованных сущностей (англ. named entity linking), могут быть учтены не только наименования объектов, но и их взаимосвязи, также представленные в тексте [*Akbik* et al., 2018].

Цель исследований – оценить качество работы сервисов, реализующих функции извлечения именованных сущностей для русского языка в виде названий и описаний географических объектов, на примере текстов биографий и фрагментов произведений сибирских писателей и предложить способы автоматизации процессов создания подобных тематических цифровых карт на основе результатов геокодирования.

В процессе исследования необходимо было решить следующие задачи:

– выбрать количественный способ оценки качества работы алгоритмов, выполняющих задачу поиска географических наименований в тексте;

– сформировать и разметить сводные таблицы данных для дальнейшего анализа алгоритмов;

– построить конвейер обработки значений, извлеченных из текста для дальнейшего геокодирования и нанесения на карту;

– подготовить карту на основе сформированных таблиц объектов.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

В последние 2–3 года появилось несколько исследований, связанных с использованием NLP-технологий в картографии и геоинформатике. Ramalho и др. [*Ramalho* et al., 2020] предлагают подход, при котором из постов в социальных сетях автоматически выбирается описание проблемы на территории города. Предлагаемое решение использует методы обработки естественного языка для геокодирования и классификации выявленных жалоб, и публикует результаты в социальной сети Crowd4City, которая аккумулирует данные о городских проблемах.

Oliveira и др. [*De Oliveira* et al., 2017] предлагают стандарт корпуса текстов на английском языке, составленного на основе сообщений из социальной сети Twitter, связанных с городскими проблемами и включающими географическую информацию. Такой набор данных может быть очень полезен для улучшения сервисов обработки текстов и разработки классификаторов при решении задач обнаружения городских проблем на основе постов из социальных сетей. Camelin и др. [*Camelin* et al., 2018] составили корпус различных новостей телевещания с французских каналов и статей в онлайн-прессе FrNewsLink, которые были вручную аннотированы, чтобы получить разметку по различным темам и связывающие аннотации между сегментами темы и статьями в прессе.

В автоматизированных системах обработки естественного языка поиск географических наименований и описаний объектов в обычных текстах относится к разделу извлечения именованных сущностей (англ. named entity recognition, NER). Именованная сущность – это одно или несколько слов, обозначающих предмет или явление заранее предопределенной категории [Кукарцев *и др.,* 2019; *Gong* et al., 2018; *Lally* et al., 2017].

Для решения поставленных задач требуется работать с двумя категориями местоположения (географического названия), которые выделяются в большинстве NER-подходов – LOC и GPE. Под данными категориями подразумеваются топонимы наименования населенных пунктов и административных единиц (GPE) и природных объектов (LOC) [*Исаченко*, 2019]. Кроме этого в тех исследованиях, которые ори-

ентированы на использование геоинформационных технологий, выделяют дополнительные подкатегории:

Bodiesofwater – объекты гидрографии,
Island – острова и полуострова,
Mountain – элементы горных систем,
Park – парки,
Transit – пути сообщения [*Ding* et al., 2021; *Honnibal, Johnson*, 2015] (рис. 1).

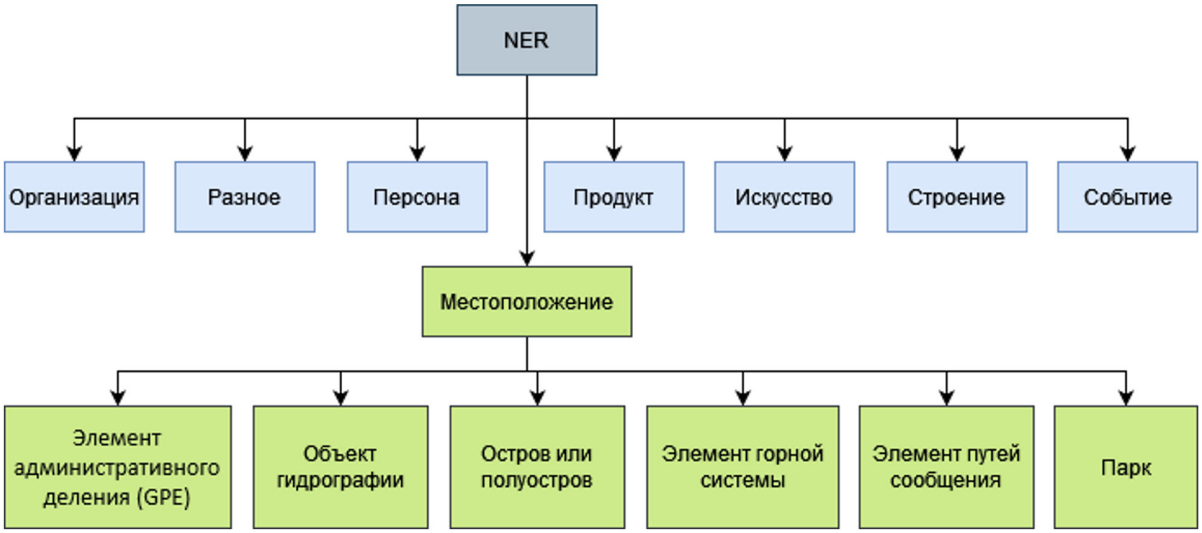

Рис. 1. Схема элементов классификации NER с расширенным обозначением географических объектов
Fig. 1. Scheme of elements of the NER classification with extended designation of geographic features

Для разметки текстов при обучении или дообучении математических моделей обработки текстов средствами машинного обучения используются схемы BIO или BILOU. Эти аббревиатуры составлены из первых букв ключевых слов, используемых для обозначения начала (B, begin), средней части (I, inside), конца (L, last) именованной сущности, кроме этого используется маркировка неиспользуемых слов (O, outside) и элементов, состоящих только из одного слова (U, unit) [*Konkol* et al., 2015]. Пример разметки словосочетания «Писатели, проживающие в Новосибирской области» для указанных схем приведен в таблице 1.

Табл. 1. Пример разметки текста для схем BIO и BILOU
Table 1. Example text markup for BIO and BILOU schemes

Слова	BIO	BILOU
Писатели	B-Person	U-Person
проживающие	O	O
в	O	O
Новосибирской	B-Location	B-Location
области	I-Location	L-Location

В качестве основного источника данных для карты были взяты тексты на русском языке, содержащие биографии и фрагменты произведений сибирских писателей с сайта Новосибирский краеведческий портал[1]. Для количественной оценки качества автоматического определения наименований географических объектов были составлены таблицы, в которых отражались описанные в тексте перемещения писателей в процессе всей

[1] Новосибирский краеведческий портал. Электронный ресурс: http://kraeved.ngonb.ru

жизни. Было проанализировано 11 писателей, среднее число географических объектов в биографии составило 12 единиц.

Сервисы, используемые в процессе работы, позволяют автоматически извлекать из текста именованные сущности и отмечать их в тексте специальными кодами. После обработки вручную были расставлены оценки качества у найденных элементов (в сравнении с теми же объектами, ранее определенными вручную). Возможные значения оценок были приняты следующие:

– - 1 – слово/словосочетание ошибочно отнесено к категории геообъектов;

– 0 – этот объект не найден этим сервисом;

– 1 – найден, но не отнесен к категориям геообъектов (LOC/GPE);

– 2 – найден частично (определены не все связанные слова или были включены ненужные);

– 3 – найден полностью.

На основе предложенных вариантов оценивания авторами предлагаются модифицированные количественные «строгие» (1–4) и «мягкие» (5–8) оценки качества работы программного обеспечения для извлечения именованных сущностей:

$$Accuracy_{strict} = \frac{c_3 + cw}{c_3 + c_w + c_0 + c_1 + c_2 + c_{-1}} \quad (1)$$

$$Precision_{strict} = \frac{c_3}{c_3 + c_2 + c_{-1}} \quad (2)$$

$$Recall_{strict} = \frac{c_3}{c_3 + c_0 + c_1} \quad (3)$$

$$F1_{strict} = 2 * \frac{Precision_{strict} * Recall_{strict}}{Precision_{strict} + Recall_{Strict}} \quad (4)$$

$$Accuracy_{soft} = \frac{c_2 + c_3 + c_w}{c_3 + c_w + c_0 + c_1 + c_2 + c_{-1}} \quad (5)$$

$$Precision_{soft} = \frac{c_2 + c_3}{c_3 + c_2 + c_{-1}} \quad (6)$$

$$Recall_{soft} = \frac{c_2 + c_3}{c_3 + c_0 + c_1} \quad (7)$$

$$F1_{soft} = 2 * \frac{Precision_{soft} * Recall_{soft}}{Precision_{soft} + Recall_{soft}} \quad (8)$$

Где, c_{-1} – количество ошибочно найденных геообъектов (ложноположительный вариант), c_0 – количество пропущенных геообъектов, c_1 – количество геообъектов, отнесенных к другому классу NER, c_2 – количество частично найденных геообъектов, c_3 – количество полностью корректно найденных геообъектов, c_w – общее число слов в тексте за исключением геообъектов. Предлагаемый способ оценки качества позволяет более гибко (с точки зрения наличия и количества ошибок частичного распознавания) оценивать наименования географических объектов, состоящих из нескольких слов, по сравнению со стандартным F1-score.

Для расчета оценки вручную была составлена таблица с правильно выделенными географическими объектами и результатами работы NER сервисов. Состав полей составляемой таблицы: ФИО писателя, наименование географического объекта, дата (день и/или месяц и/или год) начала периода местонахождения в этом месте, дата (день и/или месяц и/или год) конца периода местонахождения в этом месте, широта места, долгота места, значение типа слова по используемому сервису (сервисам) NER.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Для апробации были использованы: нейронная сеть на основе архитектуры BERT (в реализации DeepPavlov, обученной на наборе данных OntoNotes[1]), нейронная сеть на основе архитектуры Transition Based (в реализации SpaCy, обученной на наборе данных ru_core_news[2]), алгоритм на основе условных случайных полей (Conditional Random Field, в реализации Ner-RU[3]) [Белецкая, *Гриневич,* 2018], алгоритм на основе правил (в реализации Pullenti[4]). Итоговая схема исследования представлена рис. 2.

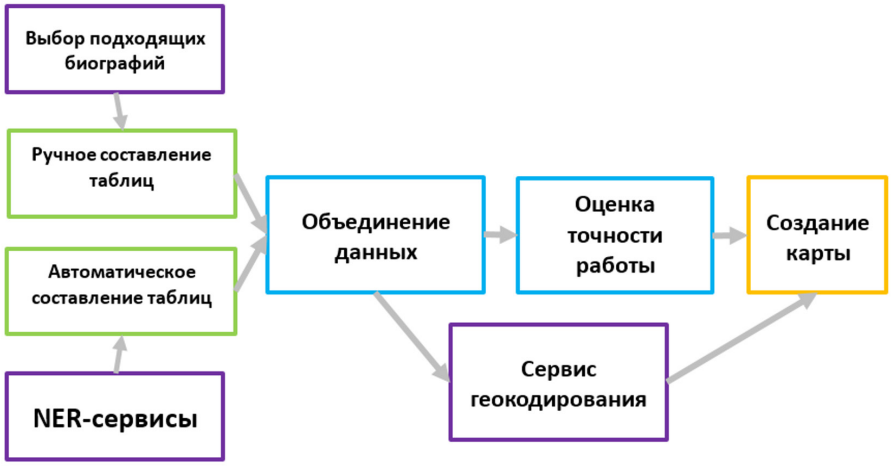

Рис. 2. Схема блоков исследования
Fig. 2. Research block diagram

Пример заполненной таблицы с ручным внесением значений географических объектов и результатами из определения с помощью NER-сервисов приведен в табл. 2. Также в таблице приведены координаты объектов на основе сервисов геокодирования Nominatim и Яндекс.Геокодер.

Табл. 2. Фрагмент данных для расчета качества работы сервисов
Table 2. Fragment of the data for calculating the quality of services

ФИО	Название географического объекта	Примечание	Широта	Долгота	DeepPavlov.ai	ner-ru.apphb.com	pullenti.ru
Вяткин Георгий Андреевич	Омская казачья станица (Омск)	Рождение	55.1227	73.3784	3	2	0
	Томск	Переезд	56.5336	84.9884	3	3	3
	Москва	Путешествие	55.5842	37.3855	3	3	3
	Санкт-Петербург	Путешествие	59.9180	30.3049	3	3	3
	Крым	Путешествие	45.2269	34.5261	3	3	3
	Финляндия	Путешествие	65.3159	25.3517	3	3	3
	Алтай	Путешествие, написание очерков о природе, жителях, их обычаях и нравах	50.9001	86.8957	3	3	3
	село Анос	В гостях у алтайского художника Г. И. Чорос-Гуркина	51.4994	85.9512	3	3	3

[1] Электронный ресурс: https://github.com/deepmipt/DeepPavlov/blob/0.17.3/deeppavlov/configs/ner/ner_ontonotes_bert_torch.json
[2] Электронный ресурс: https://github.com/explosion/spacy-models/releases/tag/ru_core_news_lg-3.3.0
[3] Электронный ресурс: https://github.com/zamgi/lingvo--Ner-ru
[4] Электронный ресурс: https://www.pullenti.ru/

ФИО	Название географического объекта	Примечание	Широта	Долгота	DeepPavlov.ai	ner-ru.apphb.com	pullenti.ru
	Харьков	Работа в газете «Утро»	49.9916	36.2805	3	3	3
	Томск	Демобилизация как учителя	56.5336	84.9884	3	3	3
	Омск	Демобилизация как учителя	55.1227	73.3784	3	3	3
	Иркутск	Эвакуация с правительственными учреждениями	52.3156	104.265	3	3	3
	Омск	Арест по доносу и доставка в Омск	55.1227	73.3784	3	3	3
	Новониколаевск (Новосибирск)	Работа в краевых журналах и газетах, в том числе в «Сибирских огнях»	55.0020	82.9560		3	3
	Вяткин					-1	
	Русском					-1	
	Ермаковом					-1	

На основе приведенных выше формул расчета качества по каждому из используемых сервисов были получены следующие итоговые значения качества распознавания именованных сущностей (строгие и мягкие соответственно):

– DeepPavlov.ai: 0.97, 1.0;
– SpaCy: 0.931, 0.986;
– ner-ru.apphb.com: 0.82, 0.903;
– pullenti.ru: 0,861, 0.963.

На текущий момент, по результатам оценки качества был выбран вариант реализации DeepPavlov.ai с наилучшими показателями. После извлечения наименований геообъектов они были геокодированы с помощью API сервиса Nominatim, и выполнена перекрестная проверка с результатами Яндекс.Геокодер. Результатом проделанной работы была визуализация данных с помощью набора слоев точечных объектов в QGIS (рис. 3).

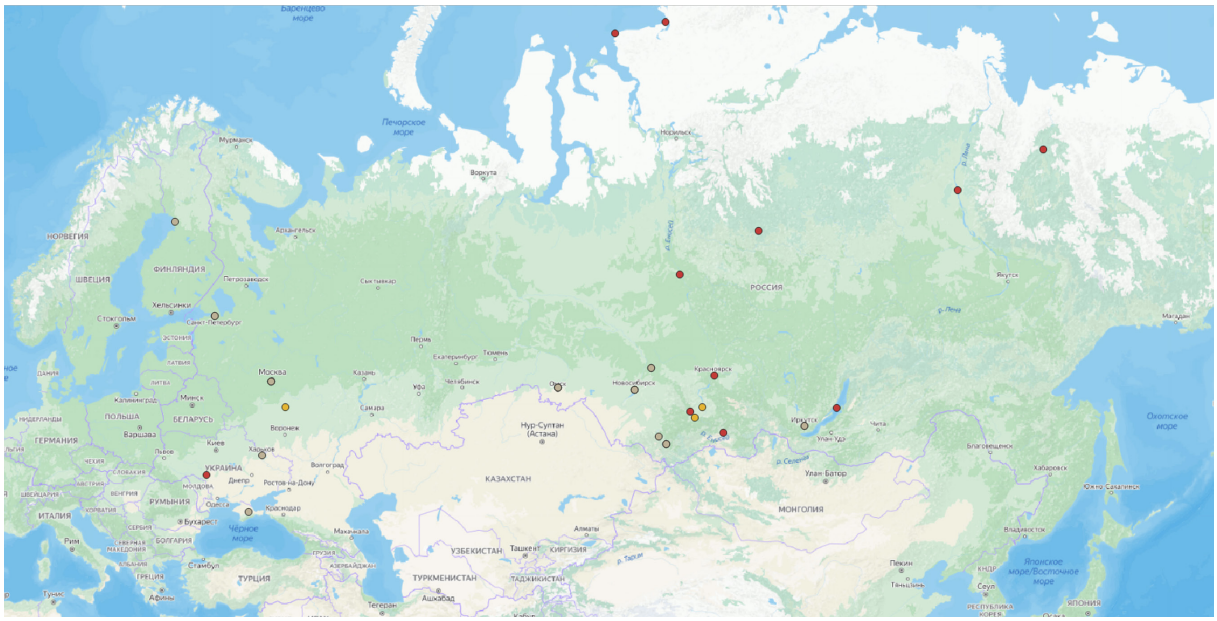

Рис. 3. Пример расположения географических объектов, связанных с биографией Кравкова М.А.
Fig. 3. An example of the location of geographical objects associated with the biography of Kravkov M.A.

ВЫВОДЫ

Поскольку все проанализированные варианты реализации используют общедоступные API и модули на языке Python в качестве основного языка разработки, то возможна автоматизация процесса построения подобных карт или геоинформационных моделей, используя средства графического программирования в ГИС или какие-либо менеджеры процессов, например, SnakeMake или AirFlow.

Для увеличения качества распознавания географических названий требуется рассмотреть и внедрить в текущее решение технологию связывания именованных сущностей (англ. named entity linking).

Итоговый вариант web-карты на основе полученных данных геокодирования предполагается реализовать с использованием модуля QGIS2Web, это потребует подготовки векторных условных знаков и организации структуры слоев, которая поможет пользователям ориентироваться в имеющихся в базе данных о писателях и их произведениях (текущий вариант web-карты размещен по адресу https://alexeykw.github.io/Siberian_writers/).

СПИСОК ЛИТЕРАТУРЫ

1. *Белецкая С.Ю., Гриневич Я.С.* Применение скрытых марковских моделей и условных случайных полей для распознавания именованных сущностей. В сборнике: Интеллектуальные информационные системы. Труды Международной научно-практической конференции. Воронеж: ВГТУ, в 2-х частях. 2018. С. 121–125.

2. *Исаченко В.В.* Обзор систем обработки текстов на естественном языке с использованием методов выделения именованных сущностей. Наука и мир, 2019. № 7-1 (71). С. 33–35.

3. *Карпачевский А.М., Филиппова О.Г.* Возможности картографирования аварийности энергосистем на основе открытых данных. ИнтерКарто. ИнтерГИС. Материалы Международной конференции, 2018. Т. 24. № 1. С. 202–211. DOI: 10.24057/2414-9179-2018-1-24-202-211.

4. *Кукарцев В.В., Колмакова З.А., Мельникова О.Л.* Системный анализ возможностей по извлечению именованных сущностей с применением технологии text mining. Перспективы науки, 2019. № 9 (120). С.18–20.

5. *Akbik A., Blythe D., Vollgraf R.* Contextual string embeddings for sequence labeling. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018. P. 1638–1649.

6. *Anh L.T., Arkhipov M.Y., Burtsev M.S.* Application of a hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition. Artificial Intelligence and Natural Language. AINL, 2017. P. 91–103. DOI: 10.1007/978-3-319-71746-3_8.

7. *Berant J., Chou A., Frostig R., Liang P.* Semantic parsing on freebase from question-answer pairs. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP). Grand Hyatt Seattle, Seattle, Washington, USA: Association for Computational Linguistics, 2013. P. 1533–1544.

8. *Bodenhamer D.J., Corrigan J., Harris T.M.* Deep maps and spatial narratives. Bloomington: Indiana University Press, 2015. 254 p. DOI: 10.2307/j.ctt1zxxzr2.

9. *Camelin N., Damnati G., Bouchekif A., Landeau A., Charlet D., Estève Y.* FrNewsLink: a corpus linking TV Broadcast News Segments and Press Articles. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), 2018. L18–1329.

10. *Cooper D., Donaldson C., Murrieta-Flores P.* Literary Mapping in the digital age. Digital research in the arts and humanities. Abingdon: Routledge, 2016. 326 p. DOI: 10.4324/9781315592596.

11. *Cura R., Dumenieu B., Abadie N., Costes B., Perret J., Gribaudi M.* Historical collaborative geocoding. ISPRS International Journal of Geo-Information, 2018. V. 7. No. 7. P. 262. DOI: 10.3390/ijgi7070262.

12. *De Oliveira M.G., De Souza Baptista C., Campelo C.E.C., Bertolotto M.* A Gold-standard Social Media Corpus for Urban Issues. Proceedings of the Symposium on Applied Computing, 2017. P. 1011–1016. DOI: 10.1145/3019612.3019808.
13. *Ding J., Wang Y., Hu W., Shi L., Qu Y.* Answering Multiple-Choice Questions in Geographical Gaokao with a Concept Graph. The semantic web – Proceedings of the 15th international conference, 2018. P. 161–176. DOI: 10.1007/978-3-319-93417-4_11.
14. *Ding N., Xu G., Chen Y., Wang X., Han X., Xie P., Zheng H., Liu Z.* Few-NERD: A Few-shot Named Entity Recognition Dataset. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021. V. 1. P. 3198–3213. DOI: 10.18653/v1/2021.acl-long.248.
15. *Gong Y., Luo H., Zhang J.* Natural Language Inference over Interaction Space. Proceedings of the 6th international conference on learning representations (ICLR), 2018.
16. *Honnibal M., Johnson M.* An Improved Non-Monotonic Transition System for Dependency Parsing. Proceedings of the 2015 Conference an Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015. P. 1373–1378. DOI: 10.18653/v1/D15-1162.
17. *Konkol, M., Konopík, M.* Segment Representations in Named Entity Recognition. Text, Speech, and Dialogue. TSD, 2015. P. 61–70. DOI: 10.1007/978-3-319-24033-6_7.
18. *Lally A., Bagchi S., Barborak M., Buchanan D.W., Chu-Carroll J., Ferrucci D. A., Glass M.R., Kalyanpur A., Mueller E.T., Murdock J.W., Patwardhan S., Prager J.M.* WatsonPaths: Scenario-based question answering and inference over unstructured information. AI magazine. Menlo Park: Association for the advancement of artificial intelligence, 2017. V. 38. No. 2. P. 59–76. DOI: 10.1609/aimag.v38i2.2715.
19. *Mozharova V., Loukachevitch N.* Two-stage approach in Russian named entity recognition. International FRUCT Conference on Intelligence, Social Media and Web. St. Petersburg: IEEE, 2016. DOI: 10.1109/FRUCT.2016.7584769.
20. *Ramalho R., Firmino A., Baptista C., Falcão A., De Oliveira M., De Andrade F.* Using Natural Language Processing for Extracting GeoSpatial Urban Issues Complaints from TV News, 2020. P. 229–239.

REFERENCES

1. *Akbik A., Blythe D., Vollgraf R.* Contextual string embeddings for sequence labeling. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe: Association for Computational Linguistics, 2018. P. 1638–1649.
2. *Anh L.T., Arkhipov M.Y., Burtsev M.S.* Application of a hybrid Bi-LSTM-CRF model to the task of Russian Named Entity Recognition. Artificial Intelligence and Natural Language. AINL, 2017. P. 91–103. DOI: 10.1007/978-3-319-71746-3_8.
3. *Beletskaya S.Y., Grinevich Y.S.* Application of Hidden Markov Models and Conditional Random Fields for Named Entity Recognition. Intelligent information systems. Proceedings of the International Scientific and Practical Conference. Voronezh: VSTU, 2018. P. 121–125 (in Russian).
4. *Berant J., Chou A., Frostig R., Liang P.* Semantic parsing on freebase from question-answer pairs. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP). Grand Hyatt Seattle, Seattle, Washington, USA: Association for Computational Linguistics, 2013. P. 1533–1544.
5. *Bodenhamer D.J., Corrigan J., Harris T.M.* Deep maps and spatial narratives. Bloomington: Indiana University Press, 2015. 254 p. DOI: 10.2307/j.ctt1zxxzr2.
6. *Camelin N., Damnati G., Bouchekif A., Landeau A., Charlet D., Estève Y.* FrNewsLink: a corpus linking TV Broadcast News Segments and Press Articles. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan: European Language Resources Association (ELRA), 2018. L18–1329.
7. *Cooper D., Donaldson C., Murrieta-Flores P.* Literary Mapping in the digital age. Digital research in the arts and humanities. Abingdon: Routledge, 2016. 326 p. DOI: 10.4324/9781315592596.

8. *Cura R., Dumenieu B., Abadie N., Costes B., Perret J., Gribaudi M.* Historical collaborative geocoding. ISPRS International Journal of Geo-Information, 2018. V. 7. No. 7. P. 262. DOI: 10.3390/ijgi7070262.
9. *De Oliveira M.G., De Souza Baptista C., Campelo C.E.C., Bertolotto M.* A Gold-standard Social Media Corpus for Urban Issues. Proceedings of the Symposium on Applied Computing, 2017. P. 1011–1016. DOI: 10.1145/3019612.3019808.
10. *Ding J., Wang Y., Hu W., Shi L., Qu Y.* Answering Multiple-Choice Questions in Geographical Gaokao with a Concept Graph. The semantic web – Proceedings of the 15th international conference, 2018. P. 161–176. DOI: 10.1007/978-3-319-93417-4_11.
11. *Ding N., Xu G., Chen Y., Wang X., Han X., Xie P., Zheng H., Liu Z.* Few-NERD: A Few-shot Named Entity Recognition Dataset. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, 2021. V. 1. P. 3198–3213. DOI: 10.18653/v1/2021.acl-long.248.
12. *Gong Y., Luo H., Zhang J.* Natural Language Inference over Interaction Space. Proceedings of the 6th international conference on learning representations (ICLR), 2018.
13. *Honnibal M., Johnson M.* An Improved Non-Monotonic Transition System for Dependency Parsing. Proceedings of the 2015 Conference an Empirical Methods in Natural Language Processing. Lisbon, Portugal: Association for Computational Linguistics, 2015. P. 1373–1378. DOI: 10.18653/v1/D15-1162.
14. *Isachenko V.V.* The review of processing systems of natural language texts using the methods of the selection of named entities. Science and world, 2019. V. 7-1(71). P. 33–35 (in Russian).
15. *Karpachevskiy A.M., Filippova O.G.* Opportunities of power systems' emergency mapping based on open data. InterCarto. InterGIS. Proceedings of the International Conference. Petrozavodsk: KRC RAS, 2018. V. 24. No. 1. P. 202–211. DOI: 10.24057/2414-9179-2018-1-24-202-211 (in Russian).
16. *Konkol, M., Konopík, M.* Segment Representations in Named Entity Recognition. Text, Speech, and Dialogue. TSD, 2015. P. 61–70. DOI: 10.1007/978-3-319-24033-6_7.
17. *Kukartsev V.V., Kolmakova Z.A., Melnikova O.L.* System analysis of possibilities to retrieve essentials using text mining technology. Science Prospects, 2019. V. 9 (120). P. 18–20 (in Russian).
18. *Lally A., Bagchi S., Barborak M., Buchanan D.W., Chu-Carroll J., Ferrucci D. A., Glass M.R., Kalyanpur A., Mueller E.T., Murdock J.W., Patwardhan S., Prager J.M.* WatsonPaths: Scenario-based question answering and inference over unstructured information. AI magazine. Menlo Park: Association for the advancement of artificial intelligence, 2017. V. 38. No. 2. P. 59–76. DOI: 10.1609/aimag.v38i2.2715.
19. *Mozharova V., Loukachevitch N.* Two-stage approach in Russian named entity recognition. International FRUCT Conference on Intelligence, Social Media and Web. St. Petersburg: IEEE, 2016. DOI: 10.1109/FRUCT.2016.7584769.
20. *Ramalho R., Firmino A., Baptista C., Falcão A., De Oliveira M., De Andrade F.* Using Natural Language Processing for Extracting GeoSpatial Urban Issues Complaints from TV News, 2020. P. 229–239.