УДК: 528.926:004 **DOI: 10.35595/2414-9179-2020-1-26-375-384**

А.А. Колесников[1], **П.М. Кикин**[2], **Дж. Нико**[3], **Е.В. Комиссарова**[4]

СИСТЕМЫ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА ДЛЯ ИЗВЛЕЧЕНИЯ ДАННЫХ И КАРТОГРАФИРОВАНИЯ НА ОСНОВЕ НЕСТРУКТУРИРОВАННЫХ БЛОКОВ ТЕКСТА

АННОТАЦИЯ

Современные технологии обработки естественного языка позволяют работать с текстами, не будучи специалистом в лингвистике. Использование популярных платформ обработки данных для разработки и использования лингвистических моделей предоставляет возможность внедрения их в популярные геоинформационные системы. Данная возможность позволяет значительно расширить функциональность и улучшить точность стандартных функций геокодирования. В статье приведено сравнение наиболее популярных методик и программного обеспечения, реализованного на их основе, на примере решения задачи извлечения географических названий из обычного текста. Такой вариант является расширенной версией операции геокодирования, поскольку в результате также получаются координаты интересующих точечных объектов, но при этом нет необходимости заранее из текста отдельно извлекать адреса или географические названия объектов. В компьютерной лингвистике эта задача решается методами извлечения именованных сущностей (*англ.* named entity recognition). Среди наиболее современных подходов к конечной реализации авторами статьи были выбраны алгоритмы, основанные на правилах, модели максимальной энтропии и свёрточные нейронные сети. Выбранные алгоритмы и методы оценивались не только с точки зрения точности поиска географических объектов в тексте, но и с позиции простоты доработки базовых правил или математических моделей с помощью собственных корпусов текста. В качестве исходных данных для апробации перечисленных методик и программных решений были выбраны отчёты о технологических нарушениях, авариях и инцидентах на объектах теплоэнергетического комплекса министерства энергетики Российской Федерации. Также приведено исследование по способу улучшения качества распознавания именованных сущностей на основе дообучения модели нейронной сети с использованием специализированного корпуса текстов.

КЛЮЧЕВЫЕ СЛОВА: географическое название, извлечение именованных сущностей, SpaCy, DeepPavlov, обработка естественного языка

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

[2] Санкт-Петербургский политехнический университет Петра Великого (СПбПУ), ул. Политехническая, д. 29, 195251, Санкт-Петербург, Россия; *e-mail:* **it-technologies@yandex.ru**

[3] Институт прикладной математики «Мауро Пиконе» (IAC), Национальный исследовательский совет Италии (CNR), ул. Амендола, д. 122/O, 75100, Бари, Италия; *e-mail:* **g.nico@ba.iac.cnr.it**

[4] Сибирский государственный университет геосистем и технологий (СГУГиТ), ул. Плахотного, д. 10, 630108, Новосибирск, Россия; *e-mail:* **komissarova_e@mail.ru**

Alexey A. Kolesnikov[1], **Pavel M. Kikin**[2], **Giovanni Niko**[3], **Elena V. Komissarova**[4]

NATURAL LANGUAGE PROCESSING SYSTEMS FOR DATA EXTRACTION AND MAPPING ON THE BASIS OF UNSTRUCTURED TEXT BLOCKS

ABSTRACT

Modern natural language processing technologies allow you to work with texts without being a specialist in linguistics. The use of popular data processing platforms for the development and use of linguistic models provides an opportunity to implement them in popular geographic information systems. This feature allows you to significantly expand the functionality and improve the accuracy of standard geocoding functions. The article provides a comparison of the most popular methods and software implemented on their basis, using the example of solving the problem of extracting geographical names from plain text. This option is an extended version of the geocoding operation, since the result also includes the coordinates of the point features of interest, but there is no need to separately extract the addresses or geographical names of the objects in advance from the text. In computer linguistics, this problem is solved by the methods of extracting named entities (*Eng.* named entity recognition). Among the most modern approaches to the final implementation, the authors of the article have chosen algorithms based on rules, models of maximum entropy and convolutional neural networks. The selected algorithms and methods were evaluated not only from the point of view of the accuracy of searching for geographical objects in the text, but also from the point of view of simplicity of refinement of the basic rules or mathematical models using their own text bodies. Reports on technological violations, accidents and incidents at the facilities of the heat and power complex of the Ministry of Energy of the Russian Federation were selected as the initial data for testing the abovementioned methods and software solutions. Also, a study is presented on a method for improving the quality of recognition of named entities based on additional training of a neural network model using a specialized text corpus.

KEYWORDS: geographical name, named entity recognition, SpaCy, DeepPavlov, natural language processing

ВВЕДЕНИЕ

В современных геоинформационных системах исходными данными для программного модуля геокодирования (геокодера) является текстовая информация в виде адреса или названия здания. При этом содержимое текстового блока можно разделить на два класса — относительные и абсолютные данные.

Абсолютный вариант — это текстовый блок, который содержит только текстовое описание пространственного объекта. Примерами могут быть почтовый индекс, полный почтовый адрес, название населённого пункта и т.п. Относительные данные, кроме абсолютного описания, содержат ключевые слова, которые в процессе обработки запроса

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

[2] Peter the Great St. Petersburg Polytechnic University (SPbPU), Polytechnicheskaya str., 29, 195251, St. Petersburg, Russia; *e-mail:* **it-technologies@yandex.ru**

[3] Institute for Applied Mathematics “Mauro Picone” (IAC), National Research Council of Italy (CNR), Via Amendola, 122/O, 75100, Bari, Italy; *e-mail:* **g.nico@ba.iac.cnr.it**

[4] Siberian State University of Geosystems and Technologies (SSUGT), Plakhotny str., 10, 630108, Novosibirsk, Russia; *e-mail:* **komissarova_e@mail.ru**

формируют дополнительные временные площадные объекты (ключевые слова: «рядом», «ближайшая») или линейные вектора (ключевые слова: «на пересечении», «через дорогу от») для поиска нужного местоположения. Запрашиваемое в относительном варианте местоположение не может быть определено без получения координат абсолютной части запроса [*Ding et al.,* 2018; *Fujita et al.*, 2014; *Gong et al.*, 2018].

В настоящий момент достаточно мало платформ геокодирования поддерживают такие запросы, и список возможных для использования ключевых слов строго предопределён разработчиками геокодера (и, как правило, сформирован только для английского языка) [*Cura et al.*, 2018]. Соответственно, фраза «через дорогу от Красного проспекта» типовым модулем геокодирования может быть понята неправильно, поскольку единственным контекстом в данном случае бывает только текущее местоположение устройства (либо координаты центра экрана), на котором выполняется запрос. Если же в предыдущем предложении речь шла о г. Новосибирске, то геокодер эту информацию не будет использовать [*Батуев и др.,* 2019; *Бешенцев и др.*, 2019].

Таким образом, авторами статьи была сформулирована задача анализа и сравнения современных возможностей анализа текста с точки зрения корректности выделения названий географических объектов и тем самым расширения возможностей стандартного геокодера для указания местоположения путём использования предложений на естественном языке, не опираясь на предопределённый набор ключевых слов, без необходимости строго выделять наименование искомого объекта из блока текста.

С точки зрения компьютерной лингвистики, задача поиска географических объектов в тексте относится к методам извлечения именованных сущностей (*англ.* named entity recognition). Именованная сущность — это одно или несколько слов, обозначающих предмет или явление определённой категории. Термин «именованная сущность» (*англ.* named entity) впервые был введён на конференции «Message Understanding Conference» (MUC-6) в 1996 г. [*Bodenhamer et al.,* 2015; *Cooper et al.*, 2016].

Наибольшее распространение среди методов этого типа получил следующий набор категорий:

- персона (PER) — имена, фамилии, отчества людей;
- местоположение (LOC) — топонимы, также в некоторых системах дополнительно разделяются наименования населённых пунктов и административных единиц (GPE) и природных объектов (LOC);
- организация (ORG) — названия организаций, компаний, объединений;
- даты (DATE) — различные варианты описания моментов времени;
- цены (MONEY) — описание стоимости с указанием валюты;
- разное (MISC) — в эту группу входят все прочие типы сущностей, если их более тщательное разделение не требуется для целей исследования [*Akbik et al.*, 2018; *Lally et al.,* 2017].

Для решаемой задачи интерес представляет только категория местоположения (географического названия), у которой наиболее часто встречающимися будут названия стран, городов, рек, собственные названия достопримечательностей, мест общепита, местных топонимов. Последние представляют наибольшую сложность, поскольку могут часто употребляться в текстах сообщений социальных сетей, но при этом не иметь соответствующей записи в базах данных геокодера (газеттирах). Вследствие этого для корректного решения задачи извлечения географических названий из текстов, собранных из разных источников, требуется не только обучить математическую модель, но и дополнить базу данных геокодера дополнительными топонимами [*Карпик и др.*, 2017; *Крылов и др.*, 2018; *Писарев*, *Ахмедов,* 2017].

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЙ

Общую схему работы современных систем обработки естественного языка можно представить в виде следующей схемы [*Honnibal, Johnson,* 2015], представленной на рис. 1. В ней блок Text представляет собой исходный неструктурированный текст, блок Doc — это результат обработки, содержащий классифицированные фрагменты текста. И блок nlp содержит все промежуточные шаги обработки. Самым первым шагом при практически любой компьютерной обработке текста является токенизация (*англ.* tokenizer) исходного текста, представляющая собой его разбивку на слова и служебные символы (знаки препинания, границы абзацев и т.п.) — токены. Далее идёт определение частей речи (*англ.* tagger или POS-tagger) на основе его определения и контекста. На следующем шаге происходит определение зависимостей и преобразование блоков текста в древовидную структуру (*англ.* parser или dependency parser). После выполнения такого преобразования можно произвести поиск и классифицирование именованных сущностей (*англ.* ner). При необходимости далее могут выполняться дополнительные операции, например, расширенное аннотирование или фильтрация.

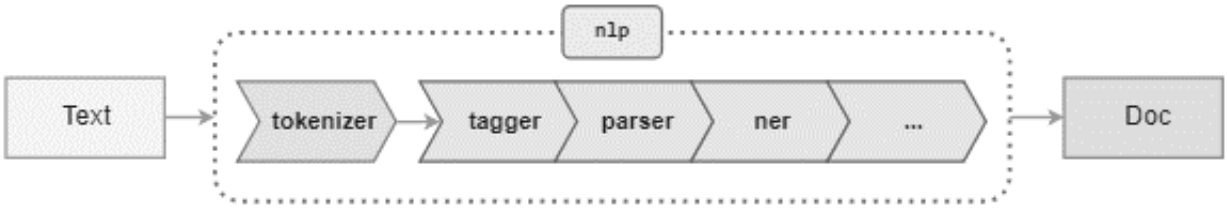

Рис. 1. Типовая схема обработки текста в системах искусственного интеллекта
Fig. 1. Typical text processing scheme in artificial intelligence systems

С точки зрения современных подходов к извлечению именованных сущностей, выделяют следующее:

- алгоритмы, основанные на правилах;
- скрытые марковские модели;
- модели максимальной энтропии;
- метод ближайших соседей;
- метод опорных векторов;
- нейронные сети (сети глубокого обучения, рекуррентные сети) [*Berant et al.,* 2013].

Если анализировать результаты сравнений и соревнований по компьютерной лингвистике, видно, что наилучшие результаты показывают технологии и программное обеспечение, использующие сочетания (ансамблирование) вышеперечисленных подходов.

Для сравнения результатов работы были выбраны несколько программных библиотек, реализующих разные подходы, являющиеся open-source и имеющие модели, работающие с русским языком:

- Natural Language Toolkit (NLTK) — используется метод максимальной энтропии в реализации Stanford CoreNLP; в качестве базовой бралась модель «Russian» [*Bird et al.,* 2009];
- Spacy — используются свёрточные нейронные сети, в качестве базовой бралась «Multi-language xx_ent_wiki_sm» модель [*Honnibal, Johnson,* 2015];

- Natasha[1] — используется алгоритм на основе правил Yargy-парсер, GLR-парсер, томита-парсер; в качестве базового набора правил брались наборы LocationExtractor и AddressExtractor [*Aycock, Horspool,* 2002]. Под парсером в данном случае понимается программное обеспечение для автоматизированного извлечения структурированных данных (фактов) из текста на естественном языке при помощи контекстно-свободных грамматик и словарей ключевых слов;
- Pullenti — алгоритм на основе правил собственного формата, POS-Tagger, в качестве искомых типов сущностей использовались Ner.GeoAnalyzer и Ner.GeoAnalyzer;
- DeepPavlov — используются глубокие нейронные сети (Bi-LSTM, двунаправленные сети долгой краткосрочной памяти и CRF, условные случайные поля); в качестве базовой бралась модель «ner_rus» [*Le et al.,* 2017; *Mozharova, Loukachevitch,* 2016].

Поскольку решаемая задача ориентирована на отдельную область именованных сущностей — географические названия, то подразумевалось, что стандартные модели программных библиотек будут недостаточно корректно распознавать требуемые слова и фразы. Исходя из этого, дополнительно рассматривались возможности по изменению либо созданию новых моделей обработки текста. Для этого в Natasha используется создание новых либо изменение существующих Yargy-правил[2] непосредственно в файлах модели, в остальных библиотеках используется метод машинного обучения, и математическая модель дообучается автоматизировано, считывая новые размеченные тексты.

В качестве исходных данных для апробации перечисленных методик и программных решений были выбраны отчёты о технологических нарушениях, авариях и инцидентах на объектах ТЭК министерства энергетики Российской Федерации, размещённые на официальном сайте[3] [*Карпачевский, Филиппова,* 2018]. Эта информация интересна тем, что представляет собой актуальные и подробные сведения о состоянии систем нефтяной, газовой, угольной промышленностей и электроэнергетики на территории всей страны. В описании аварийной ситуации представлена следующая информация: регион, объект, на котором произошла аварийная ситуация, дата и время, причины, проводимые работы по устранению, последствия. В данном исследовании использовалось и оценивалось качество извлечения первых двух категорий данных. Полученные в результате обработки текста данные впоследствии станет возможным использовать при изучении и анализе методами картографии и геоинформатики, что позволит сформировать аналитическую информацию об аварийности, произвести районирование, определить основные причины аварий и их соотношение с учётом пространственного положения и впоследствии спрогнозировать возможные чрезвычайные ситуации. С точки зрения формата, данные представляют собой сохранённые в виде растровых изображений бумажные отчёты, собранные в PDF-файлы.

Поскольку для процесса сбора описанных данных подразумевается в дальнейшем полная автоматизация, начиная от получения ссылок на файлы и до выдачи файла, содержащего точечные объекты с семантической информацией, то первым этапом работы с текстом было его распознавание с помощью какой-либо open-source OCR-библиотеки. Для распознавания был выбран Tesseract OCR, поскольку среди открытого программного обеспечения показывает лучшие результаты для русского языка и легко встраивается в

[1] Наташа — библиотека для извлечения структурированной информации из текстов на русском языке. Электронный ресурс: https://habr.com/ru/post/349864/ (дата обращения 09.12.2019)
[2] Yargy. Электронный ресурс: https://yargy.readthedocs.io/ru/latest/ (дата обращения 09.12.2019)
[3] Еженедельная информация об аварийных отключениях на объектах ТЭК, режимно-балансовая обстановка ЕЭС России. Электронный ресурс: https://minenergo.gov.ru/node/5022 (дата обращения 09.12.2019)

сторонние программные системы в виде Python модуля [*Smith*, 2007]. После распознавания из полученного текста были удалены те разделы, которые не содержали описание аварийных ситуаций (ориентируясь на ключевые слова в заголовках). В дальнейшем полученный блок текста загружался в одну из перечисленных программных библиотек. Настройки и набор элементов конвейера для предварительной обработки текста использовался, исходя из настроек библиотек по умолчанию. Для эксперимента использовались файлы от 25.07.2018 и 12.12.2018. Суммарно в этих файлах присутствовало 277 вхождений названий географических объектов.

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ И ИХ ОБСУЖДЕНИЕ

После обработки полученные результаты представляют собой массив токенов, полученных из исходного текста с тегами. Для оценки производилась выборка токенов с тегами, соответствующими местоположению (т.е. анализировались только блоки текста, помеченные ключевыми словами GPE или LOC). Оценка результатов выполнялась по F1-мере (англ. F1-score) и её элементам (precision, recall). Точность (*англ.* precision) и полнота (*англ.* recall) используются при оценке большей части алгоритмов классификации. Точность в пределах класса — это доля токенов, действительно принадлежащих данному классу, относительно всех токенов, которые модель отнесла к этому классу. Полнота — это доля найденных моделью токенов, принадлежащих классу, относительно всех токенов этого класса в тестовой выборке. F1-мера представляет собой гармоническое среднее между точностью и полнотой. Она придаёт одинаковый вес точности и полноте, поэтому F1-мера будет падать одинаково при уменьшении и точности, и полноты. Сводные результаты приведены в табл. 1.

Для получения окончательных координат использовались сервисы геокодирования Google Maps API и Here Geocoder.

Табл. 1. Оценка результатов поиска географических объектов в тексте
Table 1. Evaluation of search results for geographic features in unstructured text

	NLTK	**Natasha**	**SpaCy**	**DeepPavlov**	**Pullenti**
Precision	0	0.872	0.769	0.964	0.903
Recall	0	0.569	0.667	0.36	0.583
F1-мера	0	0.689	0.714	0.524	0.706

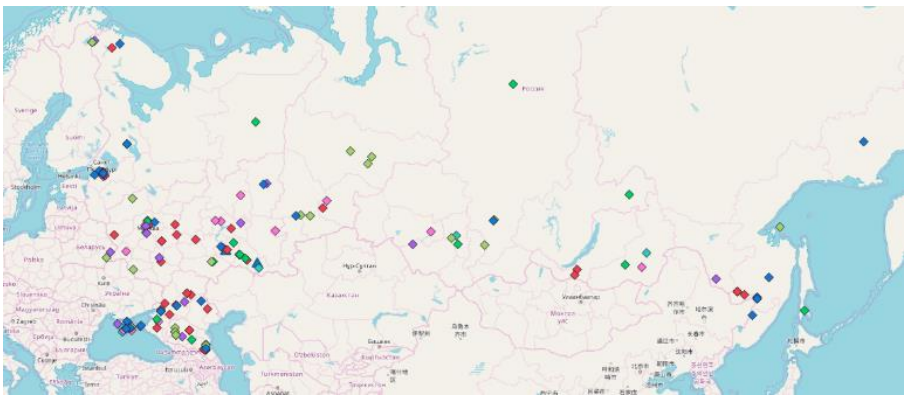

Рис. 2. Визуализация результата выделения географических сущностей (на примере SpaCy)
Fig. 2. Visualization of the result of the allocation of geographical entities (for example, SpaCy)

ВЫВОДЫ

Поскольку все проанализированные библиотеки используют Python в качестве основного языка разработки, то разработанные модули можно использовать в таких популярных геоинформационных системах, как ArcGIS, QGIS, Аксиома ГИС.

Поскольку при выделении сущностей используются стоящие рядом слова, а, например, имя региона в предложении может быть отделено несколькими промежуточными словами от названия населённого пункта, то для увеличения точности распознавания географических названий требуется рассмотреть и внедрить в текущее решение технологию связывания именованных сущностей (*англ.* named entity linking).

В результате система, подготовленная и обученная на корпусе текстов (подобранная и обработанная по определённым критериям и правилам совокупность текстов для исследования языка) определённого временного периода может хорошо справляться с другими текстами того же периода, но с течением времени результаты станут менее качественными.

СПИСОК ЛИТЕРАТУРЫ

1. *Батуев А.Р., Батуев Д.А., Бешенцев А.Н., Богданов В.Н., Дашпилов Ц.Б., Корытный Л.М., Тикунов В.С., Фёдоров Р.К.* Атласная информационная система для обеспечения социально-экономического развития Байкальского региона. ИнтерКарто. ИнтерГИС. Геоинформационное обеспечение устойчивого развития территорий: Материалы Междунар. конф. М.: Издательство Московского университета, 2019. Т. 25. Ч. 1. С. 66–80. DOI: 10.35595/2414-9179-2019-1-25-66-80.

2. *Бешенцев А.Н., Гармаев Е.Ж., Потаев В.С.* Геоинформационный мониторинг территориальных социально-экономических систем. Вестник Бурятского государственного университета. Экономика и менеджмент. Улан-Удэ: Издательство Бурятского государственного университета имени Доржи Банзарова, 2019. № 3. С. 3–9.

3. *Карпачевский А.М., Филиппова О.Г.* Возможности картографирования аварийности энергосистем на основе открытых данных. ИнтерКарто. ИнтерГИС. Материалы Международной конференции. Петрозаводск: КарНЦ РАН, 2018. Т. 24. Ч. 1. С. 202–211. DOI: 10.24057/2414-9179-2018-1-24-202-211.

4. *Карпик А.П., Лисицкий Д.В., Байков К.С., Осипов А.Г., Савиных В.Н.* Геопространственный дискурс опережающего и прорывного мышления. Вестник СГУГиТ (Сибирского государственного университета геосистем и технологий). Новосибирск: Издательство Сибирского государственного университета геосистем и технологий, 2017. Т. 22. № 4. С. 53–67.

5. *Крылов С.А., Загребин Г.И., Дворников А.В., Логинов Д.С., Фокин И.Е.* Теоретические основы автоматизации процессов атласного картографирования. Известия высших учебных заведений. Геодезия и аэрофотосъёмка. М.: Издательство Московского государственного университета геодезии и картографии, 2018. Т. 62. № 3. С. 283–293.

6. *Писарев В.С., Ахмедов Б.Н.* Автоматизированное обновление цифровых моделей геопространства. Интерэкспо ГЕО-Сибирь. Новосибирск: Издательство Сибирского государственного университета геосистем и технологий, 2017. Т. 1. № 1. С. 46–50.

7. *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.

8. *Aycock J., Horspool R.N.* Practical earley parsing. The Computer Journal, 2002. V. 4 5(6). P. 620–630. CiteSeerX 10.1.1.12.4254. DOI: 10.1093/comjnl/45.6.620.

9. *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: Association for Computational Linguistics, 2013. P. 1533–1544.
10. *Bird S., Loper E., Klein E.* Natural language processing with Python. Sebastopol, CA, USA: O'Reilly Media Inc., 2009. 512 p.
11. *Bodenhamer D.J., Corrigan J., Harris T.M.* Deep maps and spatial narratives. Bloomington: Indiana University Press, 2015. 254 p.
12. *Cooper D., Donaldson C., Murrieta-Flores P.* Literary Mapping in the digital age. Digital research in the arts and humanities. Abingdon: Routledge, 2016. 308 p.
13. *Cura R., Dumenieu B., Abadie N., Costes B., Perret J., Gribaudi M.* Historical collaborative geocoding. ISPRS International Journal of Geo-information. Basel, Switzerland: MDPI AG, 2018. V. 7. P. 262. DOI: 10.3390/ijgi7070262.
14. *Ding J., Wang Y., Hu W., Shi L., Qu Y.* Answering multiple-choice questions in geographical gaokao with a concept graph. The semantic web — 15th International conference (ESWC 2018). Heraklion, Crete, Greece. Cham: Springer, 2018. P. 161–176.
15. *Fujita A., Kameda A., Kawazoe A., Miyao Y.* Overview of Todai robot project and evaluation framework of its NLP-based problem solving. Proceedings of the 9 International conference on language resources and evaluation. Reykjavik: European Language Resources Association (ELRA), 2014. P. 2590–2597.
16. *Gong Y., Luo H., Zhang J.* Natural language inference over interaction space. 6th international conference on learning representations (ICLR). Vancouver, BC, Canada, 2018.
17. *Honnibal M., Johnson M.* An improved non-monotonic transition system for dependency parsing. Proceedings of the 2015 Conference on empirical methods in natural language processing. Lisbon, Portugal: Association for Computational Linguistics, 2015. P. 1373–1378.
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 (2). P. 59–76.
19. *Le T.A., 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. Communications in Computer and Information Science. V. 789. Cham: Springer, 2018. P. 91–103. DOI: https://doi.org/10.1007/978-3-319-71746-3_8.
20. *Mozharova V., Loukachevitch N.* Two-stage approach in Russian named entity recognition. International FRUCT conference on intelligence, social media and web (ISMW FRUCT). St. Petersburg: IEEE, 2016. DOI: 10.1109/FRUCT.2016.7584769.
21. *Smith R.* An overview of the Tesseract OCR engine. Google Inc. Proceeding 9th IEEE International conference on document analysis and recognition (ICDAR). Curitiba, Parana, Brazil: IEEE, 2007. P. 629–633.

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. *Aycock J., Horspool R.N.* Practical earley parsing. The Computer Journal, 2002. V. 45 (6). P. 620–630. CiteSeerX 10.1.1.12.4254. DOI: 10.1093/comjnl/45.6.620.
3. *Batuev A.R., Batuev D.A., Beshentsev A.N., Bogdanov V.N., Dashpilov T.B., Korytniy L.M., Tikunov V.S., Fedorov R.K.* Atlas information system for providing socio-economic development of the Baikal region. InterCarto. InterGIS. GI support of sustainable development of territories:

Proceedings of the International conference. Moscow: Moscow University Press, 2019. V. 25. Part 1. P. 66–80. DOI: 10.35595/2414-9179-2019-1-25-66-80 (in Russian, abs English).
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: Association for Computational Linguistics, 2013. P. 1533–1544.
5. *Beshentsev A.N., Garmaev E.Zh., Potaev V.S.* Geoinformation monitoring of territorial economic and social systems. Bulletin of Buryat State University. Economics and management. Ulan-Ude: Dorzhi Banzarov Buryat State University Press, 2019. V. 3. P. 3–9 (in Russian).
6. *Bird S., Loper E., Klein E.* Natural language processing with Python. Sebastopol, CA, USA: O'Reilly Media Inc., 2009. 512 p.
7. *Bodenhamer D.J., Corrigan J., Harris T.M.* Deep maps and spatial narratives. Bloomington: Indiana University Press, 2015. 254 p.
8. *Cooper D., Donaldson C., Murrieta-Flores P.* Literary Mapping in the digital age. Digital research in the arts and humanities. Abingdon: Routledge, 2016. 308 p.
9. *Cura R., Dumenieu B., Abadie N., Costes B., Perret J., Gribaudi M.* Historical collaborative geocoding. ISPRS international journal of geo-information. Basel, Switzerland: MDPI AG, 2018. V. 7. P. 262. DOI: 10.3390/ijgi7070262.
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 — 15th International conference (ESWC 2018), Heraklion, Crete, Greece. Cham: Springer, 2018. P. 161–176.
11. *Fujita A., Kameda A., Kawazoe A., Miyao Y.* Overview of Todai robot project and evaluation framework of its NLP-based problem solving. Proceedings of the 9 International conference on language resources and evaluation. Reykjavik: European Language Resources Association (ELRA), 2014. P. 2590–2597.
12. *Gong Y., Luo H., Zhang J.* Natural language inference over interaction space. 6th International conference on learning representations (ICLR). Vancouver, BC, Canada, 2018.
13. *Honnibal M., Johnson M.* An improved non-monotonic transition system for dependency parsing. Proceedings of the 2015 Conference on empirical methods in natural language processing. Lisbon, Portugal: Association for Computational Linguistics, 2015. P. 1373–1378.
14. *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. Part 1. P. 202–211. DOI: http://doi.org/10.24057/2414-9179-2018 (in Russian, abs English).
15. *Karpik A.P., Lisitsky D.V., Baykov K.S., Osipov A.G., Savinykh V.N.* Geospacial discourse of forward-looking and breaking-through way of thinking. Vestnik of the Siberian State University of Geosystems and Technologies (SSUGT). Novosibirsk: Siberian State University of Geosystems and Technologies, 2017. V. 22. No 4. P. 53–67 (in Russian).
16. *Krylov S.A., Zagrebin G.I., Dvornikov A.V., Loginov D.S., Fokin I.E.* Theoretical basics of the automatization of atlas mapping processes. Proceedings of the Higher Educational Institutions. Izvestia vuzov "Geodesy and aerophotosurveying". Moscow: Moscow State University of Geodesy and Cartography, 2018. V. 62. No 3. P. 283–293. DOI: 10.30533/0536-101X-2018-62-3-283-293 (in Russian).
17. *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 (2). P. 59–76.
18. *Le T.A., 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. Communications in Computer and Information Science. V. 789. Cham: Springer, 2018. P. 91–103. DOI: https://doi.org/10.1007/978-3-319-71746-3_8.
19. *Mozharova V., Loukachevitch N.* Two-stage approach in russian named entity recognition. International FRUCT conference on intelligence, social media and web (ISMW FRUCT). St. Petersburg: IEEE, 2016. DOI: 10.1109/FRUCT.2016.7584769.
20. *Pisarev V.S., Akhmedov B.N.* Automatic updating of digital geospace models. Proceedings of the Interexpo GEO-Sibir'. Novosibirsk: Siberian State University of Geosystems and Technologies, 2017. V. 1. No 1. P. 46–50 (in Russian).
21. *Smith R.* An overview of the Tesseract OCR engine. Google Inc. Proceeding 9th IEEE International conference on document analysis and recognition (ICDAR). Curitiba, Parana, Brazil: IEEE, 2007. P. 629–633.