DOI: 10.17323/2587-814X.2021.3.35.47

Методический подход к выявлению угрозы извлечения конфиденциальных данных из автоматизированных систем управления на базе интернет-технологий*

В.Н. Кузьмин
E-mail: vka@mil.ru

А.Б. Менисов
E-mail: vka@mil.ru

Военно-космическая академия имени А.Ф. Можайского
Адрес: 197198, г. Санкт-Петербург, ул. Ждановская, д. 13

Аннотация

В современных условиях всеохватывающей глобальной цифровизации стремительно растет и развивается киберпреступность. Поэтому обеспечение информационной безопасности относят к стратегическим целям развития информационного общества в России. Однако вопрос о том, как должно быть достигнуто «состояние защищенности личности, общества и государства от внутренних и внешних информационных угроз» в соответствии с содержанием и требованиями программ «Информационная безопасность» и «Цифровая экономика России 2024», остается открытым. Целью исследования, некоторые результаты которого представлены в настоящей публикации, является повышение эффективности выявления угроз извлечения конфиденциальных данных из автоматизированных систем управления с html-страниц для снижения риска использования этих данных на подготовительных и начальных этапах организации атак на информационную инфраструктуру государственных и коммерческих организаций. В статье описан разработанный подход к выявлению сущностей конфиденциальных данных, основанный на объединении нескольких нейросетевых технологий – универсального кодировщика предложений и нейросетевой рекуррентной архитектуры двунаправленной долгой краткосрочной памяти. Результаты оценивания показателей эффективности в сравнении с современным инструментарием обработки текстов естественного языка (spaCy) показали достоинства и перспективы практического применения данного методического подхода.

Ключевые слова: защита информации; парирование угроз информационной безопасности; конфиденциальные данные; персональные данные; машинное обучение; глубокое обучение; выявление сущностей текстов естественного языка.

Цитирование: Кузьмин В.Н., Менисов А.Б. Методический подход к выявлению угрозы извлечения конфиденциальных данных из автоматизированных систем управления на базе интернет-технологий // Бизнес-информатика. 2021. Т. 15. № 3. С. 35–47. DOI: 10.17323/2587-814X.2021.3.35.47

* Статья опубликована при поддержке Программы НИУ ВШЭ «Университетское партнерство»

Введение

Современный этап развития общества характеризуется возрастающей ролью информационной сферы, представляющей собой совокупность информации, информационной инфраструктуры, субъектов, осуществляющих сбор, формирование, распространение и использование информации, а также системы регулирования возникающих при этом общественных отношений. Информационная сфера, являясь системообразующим фактором жизни общества, активно влияет на состояние политической, экономической, оборонной и других составляющих национальной безопасности [1]. В свою очередь, развитие информационных технологий привело к преобразованию понимания личного пространства и частной жизни. Процессы, ранее происходившие в физическом (реальном) мире, перетекли в онлайн-среду: электронная торговля, поисковые сервисы, социальные сети, распространение планшетов и смартфонов, дающих людям возможность постоянно находиться в режиме онлайн. Вследствие этого объемы конфиденциальных сведений, которые человек раскрывает и выкладывает в глобальную сеть, как и персональных данных граждан, собираемых и систематизируемых разными учреждениями и ведомствами, многократно увеличились [2].

Таким образом, одной из значимых угроз национальной безопасности и интересам Российской Федерации в информационной сфере [1] является возможность использования конфиденциальной информации [3–5] на подготовительных и начальных этапах организации и проведения атак на инфраструктуру государственных и коммерческих организаций [6].

Вместе с тем, модели, методы, технологии и технические средства выявления и удаления конфиденциальных данных из открытых источников являются недостаточно совершенными из-за низкой результативности использования методов синтактического сопоставления данных, а также отсутствия глобального охвата открытого сегмента информации [7, 8]. В рамках настоящего исследования проблемная ситуация сформулирована как необходимость обеспечения эффективного выявления конфиденциальных и персональных данных с html-страниц на основе развития и реализации моделей, методов и технических средств выявления и удаления конфиденциальных данных из открытых информационных источников.

1. Анализ существующих исследований

Повышение эффективности выявления угроз утечки конфиденциальных данных может быть достигнуто посредством реализации ряда направлений деятельности, к которым относятся [9]:

- ✦ совершенствование средств мониторинга открытых источников;
- ✦ лингвистическая обработка неструктурированных данных [10];
- ✦ выявление в текстах упоминаний персон и связанных конфиденциальных данных.

В 1990-х годах выявление в текстах упоминаний впервые было представлено как задача извлечения информации (named entity recognition, NER) [11], и с тех пор данный подход привлекает большое внимание исследователей. Основная цель NER – пометить или классифицировать объекты (слова) в определенном тексте на основе заранее определенных ярлыков или тегов (например, человека, местоположения, организации и т.д.) [12]. Большая часть исследований относится к обработке английского текста, но вне зависимости от языка можно выделить три основных подхода: на основе лингвистических правил, на основе алгоритмов машинного обучения и гибридные подходы.

Лингвистический подход использует модели, основанные на правилах, которые вручную написаны лингвистами. При таком подходе формируется набор правил или шаблонов для того, чтобы различать упоминание в определенном месте текста. Разработано несколько автоматизированных систем [13, 14], в которых используются специализированные словари, включающие названия стран, крупных городов, организаций, имен людей и т.д. Основным недостатком данного подхода является необходимость использования большого объема грамматических правил в дополнение к измененному использованию (стиль, жаргон). Кроме того, эти системы практически непригодны для работы с другими языками.

Подходы, основанные на алгоритмах машинного обучения, используют большой объем аннотированных обучающих данных для получения языковых знаний высокого уровня и подразделяются на два типа: осуществляющие обучение с учителем и без учителя. Модели NER, осуществляющие обучение без учителя, не требуют никаких обучающих данных [15] и имеют возможность самостоятельного аннотирования данных. Эти модели обучения не

пользуются популярностью в практическом применении из-за достаточно низкой точности выявления сущностей в тексте. С другой стороны, модели, осуществляющие обучение с учителем, требуют большого объема качественных проаннотированных данных. Некоторые алгоритмы машинного обучения, используемых для NER (марковские модели [16–19], метод максимальной энтропии [20–22], деревья решений [23–25], метод опорных векторов [26, 27] и др.) показали надежные результаты для выявления сущностей в мультиязычных текстах.

Глубокое обучение – это подраздел машинного обучения, который представляет собой комбинацию нескольких уровней обработки данных на основе представления на нескольких уровнях абстракции. Методы глубокого обучения в последнее время широко используются из-за их выдающейся производительности по сравнению с другими методами для решения различных задач, включая обработку естественного языка.

Существуют две основные архитектуры, которые широко используются для извлечения текстового представления на уровне символов или слов [28]:

- ✦ модели на основе сверточных нейронных сетей (convolutional neural networks, CNN) [29];
- ✦ модели на основе рекуррентных нейронных сетей (recurrent neural networks RNN) [30].

Более современные архитектуры нейронных сетей, основанные на различных комбинациях сверточных и рекуррентных сетей, показывают самые высокие результаты при решении многих задач [31–34]. Эти модели демонстрируют значительную универсальность, поскольку они могут применяться к нескольким языкам с унифицированной сетевой архитектурой.

Анализ исследований в данной областей показал достоинства алгоритмов глубокого обучения для решения задачи выявления конфиденциальных данных с html-страниц.

2. Методы

В настоящем исследовании предложен методический подход, включающий применение двух разных нейросетевых технологий:

- ✦ нейросетевой рекуррентной архитектуры двунаправленной долгой краткосрочной памяти (bidirectional long short-term memory, BLSTM), которая показала улучшение качества выявления сущностей для решения других задач [35, 36];
- ✦ универсального кодировщика предложений, позволяющего масштабировать подход для текстов разных языков [37].

Объединение двунаправленной нейронной сети долгой краткосрочной памяти (BLSTM) и кодировщика предложений включает в себя следующие этапы (*рисунок 1*):

1. Очистка html-страницы от разметки и другой служебной информации. Выделение текстовой части, находящейся на html-странице.

2. Предобработка текста.

3. Представление признаков текста: преобразование первичных признаков в векторное представление.

4. Представление предложений на основе BLSTM: получение высокоуровневого представления признаков (семантики) из признаков этапа 3.

5. Построение общего вектора признаков: объединение признаков предложения лексического и семантического уровней из этапов 3 и 4 с целью формирования окончательного вектора признаков.

6. Классификация: определение конфиденциальных данных.

Рассмотрим более подробно каждый из этапов выявления конфиденциальных сущностей на html-страницах.

Этап 1. Очистка html-страницы. На данном этапе необходимо решить задачи извлечения фактического текста из содержимого html-страницы и очистки текста от специальных символов. Процесс начинается с построения объектной модели документа (document object model, DOM) html-страницы путем анализа тегов. Модель DOM обеспечивает представление структуры html-страницы. Текстовые узлы модели DOM извлекаются для дальнейшего использования: текст фильтруется и очищается, с исключением скриптов и символов структуры html, таких как списки навигации, теги стилей, таблицы и различные фреймы. На этом этапе также удаляются знаки препинания и специальные символы.

Этап 2. Предобработка текста. На этом этапе полученные текстовые сегменты токенизируются, чтобы получить отдельные токены (слова). Затем удаляются стоп-слова – часто встречающиеся слова на разных языках, которые имеют наименьшее семантическое значение. Сначала идентифицируется язык текста, а затем определяется список стоп-слов для этого языка. Это не противоречит цели предоставления не зависимого от языка метода, посколь-

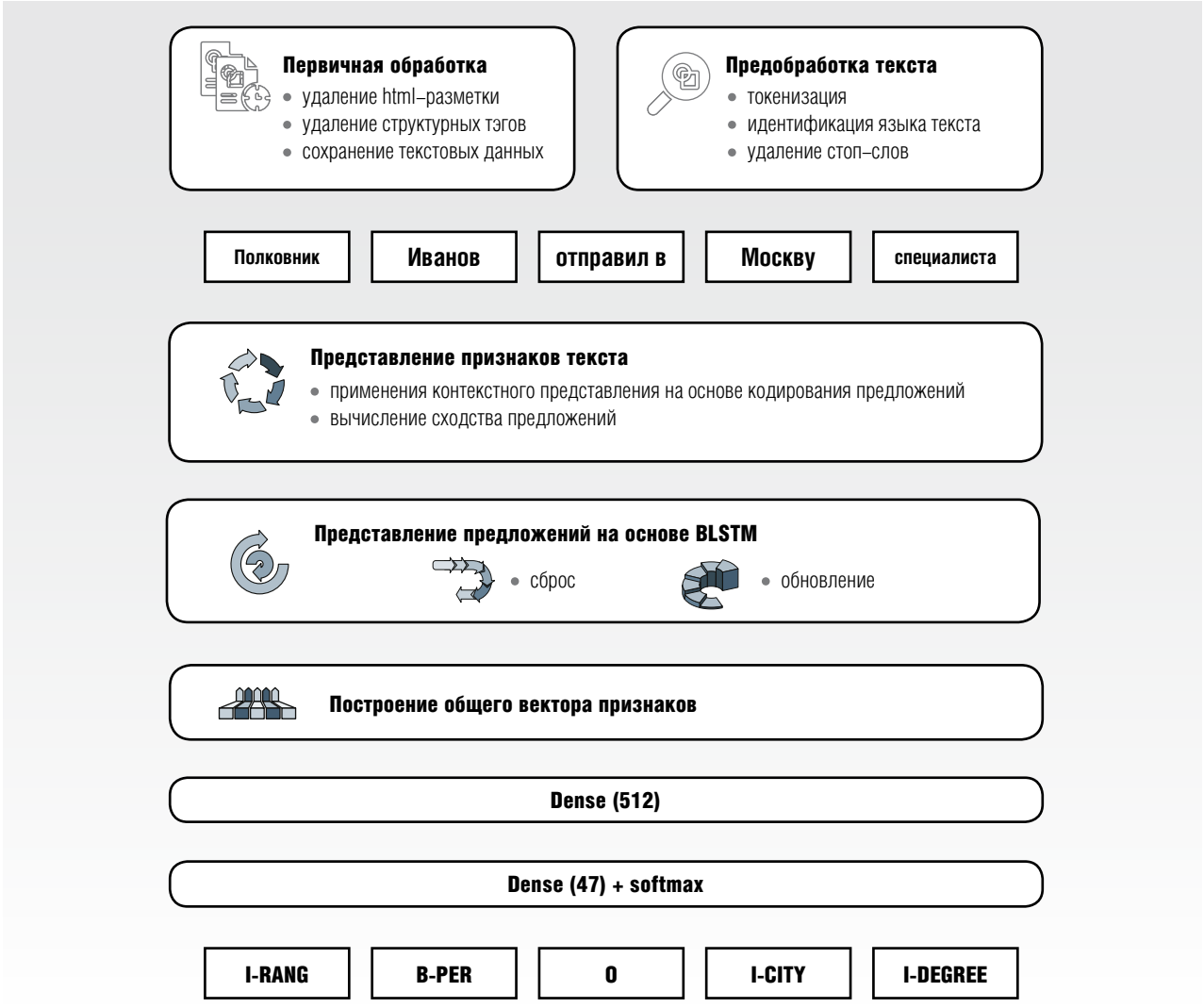

Рис. 1. Схема разработанного методического подхода извлечения конфиденциальных данных

ку средство определения языка и список стоп-слов для разных языков доступны в открытых программных библиотеках.

В сложных естественных языках (таких как русский) одно и то же слово может принимать разные формы (падежи), и в словарь частотного анализа могут попадать все словоформы, отличающиеся предлогами и окончаниями. Из-за этого может сильно увеличиваться размер словаря и, соответственно, размер набора данных для обучения, что может вызывать уменьшение производительности системы и ухудшение обобщающих способностей классификатора (переобучения). Для решения этой проблемы применяются дополнительные меры предобработки текста: лемматизация и стемминг.

Для решения задачи выявления конфиденциальных данных эти дополнительные меры были специально пропущены, что объясняется следующими причинами:

- ✦ стемминг (и особенно лемматизация) требуют словарей языков, на которых написан текст;
- ✦ эти меры выполняют поиск по словарям нормальной формы для всех слов текста, что может существенно снижать производительность;
- ✦ конфиденциальные данные часто содержат информацию, представленную в форме цифр и специальных знаков (например, номера удостоверений и других документов).

Этап 3. Представление признаков текста. Представление признаков текста – один из основных эта-

пов его лингвистической обработки. Для решения задачи выявления конфиденциальных данных предлагается применять предварительно обученный алгоритм представления предложений (или последовательностей слов) под названием Universal Sentence Encoder (USE) [37]. USE – это алгоритм кодирования предложений, выпущенный Google в 2018 году, цель которого – обеспечить представление на уровне предложений, а не на уровне слов или символов.

Алгоритм USE был реализован с использованием двух подходов:

- ✦ применения контекстного представления на основе кодирования предложений;
- ✦ оценки сходства предложений.

Алгоритм USE впервые был применен для английского языка [37], а затем был реализован для многоязычных текстов [38].

Этап 4. Представление предложений на основе BLSTM. На этом этапе используется двунаправленный рекуррентный блок как расширенная версия рекуррентной нейронной сети. Облегченная структура блока с использованием двух вентилей (сброса и обновления) позволяет повысить оперативность решения проблемы исчезновения градиента по сравнению с архитектурой долгой краткосрочной памяти (long short-term memory, LSTM) [39], которая состоит из трех вентилей (вход, выход и вентиль забывания).

Этап 5. Построение общего вектора признаков. На данном этапе происходит объединение признаков предложения лексического и семантического уровней с целью сохранения всех возможных отличительных особенностей текста. Объединение производится в конкатинационном слое.

Этап 6. Классификация. На этапе классификации вначале плотно-связанный слой из 512 нейронов используется для обработки отличительных характеристик текста из слоя конкатенации признаков. Далее используется слой с элементом активации softmax для вычисления вероятности того, что данное слово относится к одному из 57 классов. Набор данных аннотируется с использованием формата маркировки IOB [40], и включает 28 классов конфиденциальной информации (B- и I-тегов) и один дополнительный класс (O-тег). Поэтому модель позволяет проводить классификацию по 57 классам.

3. Результаты

В данном разделе описано проведение эксперимента и параметры, которые использовались для обучения и валидации предложенного подхода.

3.1. Набор данных

Набор данных, используемый для исследований, был предоставлен организацией Kaspersky Innovation Hub[1], в рамках Всероссийского хакатона Digital SuperHero (Fintech&Security) 2020 года[2]. Набор данных состоит из 28 классов конфиденциальных данных (*таблица 1*). В данной таблице также представлено распределение вхождений по классам конфиденциальных данных, которые использовались для обучения и валидации предлагаемых в подходе моделей. Всего набор данных состоит из почти из 900 html-страниц, из которых 833 html-страницы использовались в качестве обучающего набора данных, а 70 html-страниц – для валидации модели.

3.2. Метрики качества

Чтобы оценить эффективность предлагаемого подхода, использовалось несколько метрик. Сначала была измерена точность, которая является известным показателем для оценки любой модели машинного или глубокого обучения и характеризует долю правильно классифицированных объектов из их общего числа. Показатель точности вычисляется следующим образом:

$$Pr = \frac{TP}{TP + FP}, \quad (1)$$

где TR – число истинно-положительных результатов;

FP – число ошибок первого рода (ложноположительное срабатывание).

Затем была оценена метрика полноты, которая представляет собой количество правильно определенных объектов из общего числа объектов в наборе данных. Данная метрика вычисляется следующим образом:

$$R = \frac{TP}{TP + FN}, \quad (2)$$

где TP – число истинно-положительных результатов;

[1] https://www.kaspersky.ru/ihub/
[2] https://www.dshkazan.ru/finsec/

FN – число ошибок второго рода (ложноотрицательное срабатывание).

Наконец, *F*1-мера рассчитывается на основе значений точности и полноты:

$$F1 = 2\frac{Pr \cdot R}{Pr + R}. \quad (3)$$

где *Pr* – значение показателя точности;

R – значение показателя полноты.

3.3. Качество выявления конфиденциальных данных

В настоящее время на рынке средств анализа открытых источников информации и применения технологии искусственного интеллекта есть много программных решений (например, Amazon, ABBYY, IBM Watson, MS Azure и продукты компании Palantir). Однако для сравнения с разработанным подходом был принят фреймворк spaCy[3], что объясняется следующими причинами экономического и технического характера:

- ✦ поддержка более 60 языков;
- ✦ готовая к промышленному применению система обучения лингвистических моделей;
- ✦ наличие расширяемых компонент для распознавания именованных сущностей, тегирования частей речи, синтаксического анализа зависимостей, сегментации предложений, классификации текста, лемматизации, морфологического анализа, связывания сущностей и т.д.;
- ✦ поддержка пользовательских моделей на PyTorch, TensorFlow и других нейросетевых фреймворках;
- ✦ наличие встроенных визуализаторов для синтаксиса и NER;
- ✦ относительно простая интеграция модели в автоматизированные системы.

Сравнение результатов, представленных на хакатоне[4] и полученных с помощью фреймворка spaCy показало, что разработанный подход показывает повышение качества выявления конфиденциальных данных на 21% по всем классам конфиденциальных данных (*таблица 2*). Подход обеспечил достижение средней $F1 = 0{,}55$, средней точности $Pr = 0{,}57$ и средней полноты $R = 0{,}67$.

Таблица 1.

Классы использованного набора данных

№ п/п	Обозначение	Характеристика	Количество вхождений
1	PASSPORT	Паспортные данные	1
2	DRIVER_LIC	Данные водительского удостоверения	1
3	CAR_START	Начало использования автомобиля	8
4	CAR	Государственный регистрационный номер автомобиля	28
5	EDU_START	Начало обучения	35
6	DEATHDATE	Дата смерти	83
7	AGE	Возраст	120
8	EDU_END	Окончание учебы	145
9	BIRTHDATE	Дата рождения	146
10	ZIPCODE	Почтовый индекс	218
11	FACULTY	Факультет	219
12	HOBBY	Хобби	315
13	START	Время начала обучения или работы	338
14	EMAIL	Адрес электронной почты	354
15	END	Окончание работы	355
16	STREET	Улица	364
17	EDU	Образование	389
18	TEL	Телефон	409
19	DEGREE	Звание, должность	459
20	STATE	Подразделение	912
21	NICKNAME	Профиль социальной сети (логин)	1220
22	COUNTRY	Страна	1254
23	INDUSTRY	Направление деятельности	1597
24	CITY	Город	1824
25	GENDER	Пол	1960
26	FUNC	Должностные обязанности	2517
27	ORG	Название организации	3667
28	PER	ФИО (или его часть)	7682

[3] https://www.spaCy.com
[4] https://www.dshkazan.ru/finsec/

Таблица 2.

Сравнение результатов моделей для выявления конфиденциальных данных

№ п/п	Класс	Точность		Полнота		*F1*	
		Авторский подход	spaCy	Авторский подход	spaCy	Авторский подход	spaCy
1	PER	0,9082	0,35	0,87076	0,98	0,88908	0,51579
2	ORG	0,1932	0,44	0,93319	0,75	0,32008	0,55462
3	FUNC	0,1872	0,15	0,51531	0,79	0,27466	0,25213
4	CITY	0,584	0,23	0,58657	0,83	0,58528	0,36019
5	NICKNAME	0,7695	0,16	0,94219	0,88	0,84714	0,27077
6	COUNTRY	0,8086	0,45	0,51655	0,92	0,63038	0,60438
7	GENDER	0,9379	0,18	0,29631	0,06	0,45034	0,09
8	INDUSTRY	0,5283	0,3	0,2601	0,08	0,34858	0,12632
9	STATE	0,1797	0,17	0,72339	0,83	0,28783	0,2822
10	EMAIL	0,8901	0,34	0,41612	0,19	0,56712	0,24377
11	STREET	0,7692	0,25	0,69175	0,8	0,72844	0,38095
12	TEL	0,8774	0,5	0,5833	0,76	0,70075	0,60318
13	EDU	0,171	0,16	0,80821	0,8	0,28224	0,26667
14	ZIPCODE	0,6034	0,41	0,77374	0,85	0,67807	0,55318
15	DEGREE	0,6585	0,35	0,93206	0,1	0,77175	0,15556
16	START	0,3947	0,27	0,43456	0,84	0,41368	0,40865
17	EDU_END	0,1366	0,45	0,55193	0,3	0,21896	0,36
18	END	0,0298	0,38	0,86437	0,92	0,05754	0,53785
19	AGE	0,9287	0,18	0,83912	0,76	0,88164	0,29106
20	HOBBY	0,1011	0,3	0,5287	0,21	0,16975	0,24706
21	BIRTHDATE	0,8076	0,17	0,92066	0,5	0,86043	0,25373
22	FACULTY	0,9578	0,34	0,97121	0,45	0,96448	0,38734
23	CAR	0,4014	0,25	0,99413	0,85	0,57185	0,38636
24	DEATHDATE	0,6176	0,5	0,73078	0,96	0,66942	0,65753
25	EDU_START	0,9323	0,16	0,78325	0,02	0,85131	0,03556
26	CAR_START	0,3057	0,15	0,146	0,2	0,19761	0,17143
27	PASSPORT	0,8492	0,23	0,72283	0,53	0,78096	0,32079
28	DRIVER_LIC	0,5959	0,16	0,44169	0,54	0,50733	0,24686
Среднее значение:		**0,57588**	**0,285**	**0,67067**	**0,59643**	**0,55381**	**0,34157**

4. Дискуссия

Чтобы продемонстрировать практическую значимость предлагаемого методического подхода выявления конфиденциальных данных из сети интернет, произведено сравнение полученных результатов с другими композициями нейросетевых технологий (*таблица 3*).

Как видно из *таблицы 3*, модель BLSTM, примененная в разработанном подходе, превосходит модель LSTM во всех указанных классах. Можно отметить, что самое незначительное улучшение было достигнуто в классе EDU_END, в то время как самое высокое – в классе PER. Это можно объяснить распределением данных по классам и их размерам. Возвращаясь к *таблице 1*, можно заметить, что у

Таблица 3.

Сравнение результатов работы разных композиций методического подхода

№ п/п	Класс	Точность				Полнота				*F1*			
		Авторский подход	LSTM	Без USE	USE	Авторский подход	LSTM	Без USE	USE	Авторский подход	LSTM	Без USE	USE
1	PER	0,9082	0,788899	0,7554	0,2513	0,87076	0,3161	0,9247	0,4166	0,88908	0,45136	0,83153	0,31349
2	ORG	0,1932	0,910661	0,7594	0,3073	0,93319	0,0655	0,6615	0,4548	0,32008	0,12222	0,70711	0,36673
3	FUNC	0,1872	0,564068	0,2792	0,8071	0,51531	0,2536	0,0691	0,5931	0,27466	0,34992	0,11074	0,68375
4	CITY	0,584	0,979289	0,1985	0,5707	0,58657	0,4723	0,3151	0,8761	0,58528	0,63722	0,24358	0,69113
5	NICKNAME	0,7695	0,958619	0,6594	0,9251	0,94219	0,9439	0,3993	0,0077	0,84714	0,9512	0,4974	0,01521
6	COUNTRY	0,8086	0,682749	0,9073	0,8715	0,51655	0,1534	0,7579	0,4965	0,63038	0,25058	0,82588	0,63264
7	GENDER	0,9379	0,761573	0,3181	0,0151	0,29631	0,7039	0,2468	0,4464	0,45034	0,73159	0,27794	0,02926
8	INDUSTRY	0,5283	0,155593	0,2821	0,1805	0,2601	0,2655	0,8068	0,0553	0,34858	0,19621	0,41799	0,08467
9	STATE	0,1797	0,109872	0,0281	0,2822	0,72339	0,2258	0,8331	0,873	0,28783	0,14783	0,05438	0,42655
10	EMAIL	0,8901	0,712207	0,4924	0,5611	0,41612	0,3041	0,3899	0,8125	0,56712	0,42625	0,4352	0,66379
11	STREET	0,7692	0,154987	0,7198	0,9594	0,69175	0,9779	0,9269	0,7743	0,72844	0,26757	0,8103	0,85698
12	TEL	0,8774	0,036389	0,7911	0,4411	0,5833	0,1756	0,5942	0,4934	0,70075	0,06028	0,67864	0,46582
13	EDU	0,171	0,26002	0,0648	0,5232	0,80821	0,1515	0,7556	0,2679	0,28224	0,19149	0,11942	0,35438
14	ZIPCODE	0,6034	0,203505	0,9409	0,2833	0,77374	0,1067	0,5478	0,7517	0,67807	0,13998	0,69245	0,41156
15	DEGREE	0,6585	0,299845	0,3623	0,9814	0,93206	0,8859	0,242	0,0327	0,77175	0,44805	0,2902	0,06328
16	START	0,3947	0,436453	0,8491	0,716	0,43456	0,7234	0,3869	0,9749	0,41368	0,54442	0,53161	0,82562
17	EDU_END	0,1366	0,806458	0,8001	0,8182	0,55193	0,3229	0,2334	0,4665	0,21896	0,46118	0,3614	0,59419
18	END	0,0298	0,858825	0,7621	0,4577	0,86437	0,9984	0,3264	0,242	0,05754	0,92338	0,45707	0,31662
19	AGE	0,9287	0,239638	0,4984	0,4532	0,83912	0,3478	0,8059	0,8474	0,88164	0,28376	0,61586	0,59055
20	HOBBY	0,1011	0,116531	0,9497	0,9496	0,5287	0,5741	0,0152	0,4049	0,16975	0,19373	0,02987	0,56771
21	BIRTHDATE	0,8076	0,133296	0,9161	0,9204	0,92066	0,6129	0,15	0,3458	0,86043	0,21897	0,25785	0,50276
22	FACULTY	0,9578	0,995736	0,0912	0,8146	0,97121	0,0517	0,6599	0,3144	0,96448	0,09824	0,16023	0,4537
23	CAR	0,4014	0,378413	0,2613	0,821	0,99413	0,3192	0,6812	0,0832	0,57185	0,34627	0,37774	0,15108
24	DEATHDATE	0,6176	0,634855	0,9917	0,3356	0,73078	0,3214	0,8013	0,3979	0,66942	0,42675	0,88636	0,3641
25	EDU_START	0,9323	0,5145	0,4761	0,5342	0,78325	0,1459	0,0167	0,0504	0,85131	0,22736	0,03235	0,09205
26	CAR_START	0,3057	0,4658	0,8526	0,8066	0,146	0,9981	0,5294	0,7014	0,19761	0,63518	0,6532	0,75031
27	PASSPORT	0,8492	0,689894	0,3281	0,5283	0,72283	0,6247	0,3017	0,9222	0,78096	0,6557	0,31436	0,67176
28	DRIVER_LIC	0,5959	0,390814	0,511194	0,012905	0,44169	0,233	0,0786	0,6786	0,50733	0,29193	0,13618	0,02533
Среднее значение:		**0,57588**	**0,5085532**	**0,56595**	**0,57602**	**0,67067**	**0,43841**	**0,48062**	**0,4922**	**0,55381**	0,38138	0,42167	0,42732

класса PER самый большой объем по сравнению с другими классами (7682 именованных объектов), в то время как у класса EDU_END – самый маленький объем (145 именованных объектов).

Однако, для таких классов конфиденциальных данных, как ORG и COUNTRY наблюдается пониженное качество, что объясняется следующими причинами:

- ✦ объективное преимущество выявления этих классов фреймворком spaCy, поскольку возможность выявления класса ORG присутствует во всех лингвистических моделях spaCy для разных языков, обученных на большем объеме данных; также имеется возможность выявления класса COUNTRY с помощью дополнительных классов spaCy (GPE и LOC);
- ✦ пересечение сущностей разных классов в обучающей выборке: ORG, STATE, EDU и FACULTY.

К числу причин повышения качества выявления конфиденциальных данных относятся:

- ✦ возможности алгоритма USE в части обработки семантических представлений предложений и фраз. Использование только алгоритма USE в качестве классификатора для того же набора данных, как показано в *таблице 3* (USE), позволило достичь средней $F1 = 0{,}42$. Хотя общие результаты алгоритма USE для классификации ниже, чем результаты, достигнутые с помощью разработанного подхода, можно увидеть, что использование алгоритма USE только для классификации некоторых других классов (STREET) выше, чем у предложенного подхода ($F1 = 0{,}81$ у USE, против 0,73 у предложенного подхода). Следовательно, в дальнейших исследованиях следует провести дополнительную работу по улучшению классификации отдельных массивов данных;
- ✦ предложенная последовательность этапов и архитектурой нейронной сети. В частности, путем аккумуляции алгоритма USE и BLSTM было достигнуто улучшение средней F1-меры на 15%.

Заключение

Огромный объем неструктурированных данных сети интернет, распространяемый ежедневно, вызывает потребность в разработке эффективных методов поиска и извлечения информации. Извлечение конфиденциальных данных – сложная задача классификации для текстов естественного языка, которая еще более усложняется при применении к html-страницам из-за их особых свойств и сложной структуры. В настоящей статье представлен новый подход глубокого обучения для выявления конфиденциальных данных, который доказал свою эффективность по сравнению с другими.

Основная цель разработки нового подхода – предоставить более детализированные результаты для практического применения в области обработки естественного языка и информационной безопасности. В разработанном подходе использована нейросетевая технология двунаправленной долгой краткосрочной памяти в сочетании с многоязычным универсальным кодировщиком предложений. ■

Литература

1. Доктрина информационной безопасности Российской Федерации. Утверждена Указом Президента Российской Федерации от 5 декабря 2016 г. № 646. [Электронный ресурс]: http://pravo.gov.ru/proxy/ips/?docbody=&prevDoc=102161033&backlink=1&&nd=102417017 (дата обращения 01.02.2021).
2. Шайдуллина В.К. Большие данные и защита персональных данных: основные проблемы теории и практики правового регулирования // Общество: политика, экономика, право. 2019. № 1 (66). С. 51–55. DOI: 10.24158/pep.2019.1.8.
3. Федеральный закон РФ от 27.07.2006 №152-ФЗ «О персональных данных». [Электронный ресурс]: http://pravo.gov.ru/proxy/ips/?docbody&nd=102108261 (дата обращения 01.02.2021).
4. Указ Президента Российской Федерации от 06.03.1997 г. № 188 «Об утверждении перечня сведений конфиденциального характера». [Электронный ресурс]: http://pravo.gov.ru/proxy/ips/?docbody=&firstDoc=1&lastDoc=1&nd=102046005 (дата обращения 01.02.2021).
5. Федеральный закон от 27.07.2004 г. № 79-ФЗ «О государственной гражданской службе Российской Федерации». [Электронный ресурс]: http://pravo.gov.ru/proxy/ips/?docbody=&firstDoc=1&lastDoc=1&nd=102088054 (дата обращения 01.02.2021).
6. Информационное сообщение «О разработке методического документа ФСТЭК России «Методика определения угроз безопасности информации в информационных системах» от 9 апреля 2020 г. № 240/22/1534. [Электронный ресурс]: https://fstec.ru/normotvorcheskaya/informatsionnye-i-analiticheskie-materialy/2071-informatsionnoe-soobshchenie-fstek-rossii-ot-9-aprelya-2020-g-n-240-22-1534 (дата обращения 01.02.2021).

7. Лось В.П., Никульчев Е.В., Пушкин П.Ю., Русаков А.М. Информационно-аналитическая система мониторинга выполнения операторами персональных данных требований законодательства // Проблемы информационной безопасности. Компьютерные системы. 2020. № 3. С. 16–23.
8. Козин И.С. Метод обеспечения безопасности персональных данных при их обработке в информационной системе на основе анализа поведения пользователей // Информационно-управляющие системы. 2018. № 3. С. 69–78. DOI: 10.15217/issn1684-8853.2018.3.69.
9. Ивичев В.А., Игнатова Т.В. Технологии выявления и очистки персональных данных открытых источников // ЭКО. 2013. № 2 (464). С. 168–179.
10. Платонов А.А., Псарев А.А. Подход к выявлению дефектов программ на основе применения взаимодополняющих семантик // Труды Военно-космической академии имени А.Ф. Можайского. 2019. № 666. С. 148–157.
11. Grishman R. Whither written language evaluation? // Proceedings of the Workshop on Human Language Technology. Plainsboro, New Jersey, 8–11 March 1994. P. 120–125.
12. Minkov E., Wang R.C., Cohen W. Extracting personal names from email: Applying named entity recognition to informal text // Proceedings of the Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada, 6–8 October 2005. P. 443–450.
13. Grishman R. The NYU System for MUC-6 or Where's the Syntax? // Proceedings of the Sixth Message Understanding Conference (MUC-6). Columbia, Maryland, 6–8 November 1995. P. 167–175.
14. Wakao T., Gaizauskas R., Wilks Y. Evaluation of an algorithm for the recognition and classification of proper names // Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996). Copenhagen, 5–9 August 1996. Vol. 1. P. 418–423.
15. Alfred R., Leong L.C., On C.K., Anthony P. Malay named entity recognition based on rule-based approach // International Journal of Machine Learning and Computing. 2014. Vol. 4. No 3. P. 300–306. DOI: 10.7763/IJMLC.2014.V4.428.
16. Salleh M.S., Asmai S.A., Basiron H., Ahmad S. Named entity recognition using fuzzy c-means clustering method for Malay textual data analysis // Journal of Telecommunication, Electronic and Computer Engineering. 2018. Vol. 10. No 2–7. P. 121–126.
17. Zhou G.D., Su J. Named entity recognition using an HMM-based chunk tagger // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA, USA, 6–12 July 2002. P. 473–480.
18. Morwal S., Jahan N., Chopra D. Named entity recognition using hidden Markov model (HMM) // International Journal on Natural Language Computing. 2012. Vol. 1. No 4. P. 15–23. DOI: 10.5121/ijnlc.2012.1402.
19. Morwal S., Jahan N. Named entity recognition using hidden Markov model (HMM): an experimental result on Hindi, Urdu and Marathi languages // International Journal of Advanced Research in Computer Science and Software Engineering. 2013. Vol. 3. No 4. P. 671–675.
20. Borthwick A. A maximum entropy approach to named entity recognition (PhD Thesis). New York: New York University, 1999.
21. Chieu H.L., Ng H.T. Named entity recognition: a maximum entropy approach using global information // Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002). Taipei, Taiwan, 24 August – 1 September 2002. Vol. 1. P. 1–7. DOI: 10.3115/1072228.1072253.
22. Chieu H.L., Ng H.T. Named entity recognition with a maximum entropy approach // Proceedings of the Seventh Conference on Natural language learning at HLT-NAACL 2003. Edmonton Canada, 31 May 2003. P. 160–163.
23. Speck R., Ngomo A.C.N. Ensemble learning for named entity recognition // Proceedings of the 13th International Semantic Web Conference (ISWC 2014). Riva del Garda, Italy, 19–23 October 2014. P. 519–534.
24. Paliouras G., Karkaletsis V., Petasis G., Spyropoulos C.D. Learning decision trees for named-entity recognition and classification // Proceedings of the ECAI Workshop on Machine Learning for Information Extraction (ECAI 2000). Berlin, 21 August 2000. P. 1–6.
25. Szarvas G., Farkas R., Kocsor A. A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms // Proceedings of the 9th International Conference on Discovery Science (DS 2006). Barcelona, Spain, 7–10 October 2006. P. 267–278. DOI: 10.1007/11893318_27.
26. Mansouri A., Affendey L. S., Mamat A. Named entity recognition approaches // International Journal of Computer Science and Network Security. 2008. Vol. 8. No 2. P. 339–344.
27. Ekbal A., Bandyopadhyay S. Named entity recognition using support vector machine: A language independent approach // International Journal of Electrical, Computer, and Systems Engineering. 2010. Vol. 4. No 2. P. 155–170.
28. Li J., Sun A., Han J., Li C. A survey on deep learning for named entity recognition // IEEE Transactions on Knowledge and Data Engineering. 2020. (Early access). DOI: 10.1109/TKDE.2020.2981314.
29. Ma X., Hovy E. End-to-end sequence labeling via bi-directional lstm-cnns-crf // arXiv:1603.01354v5. 2016.
30. Leveraging linguistic structures for named entity recognition with bidirectional recursive neural networks / P.-H. Li [et al.] // Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 7–11 September 2017. P. 2664–2669.
31. Fu J., Liu P., Zhang Q. Rethinking generalization of neural models: A named entity recognition case study // Proceedings of the 34th AAAI Conference on Artificial Intelligence. New York, USA, 7–12 February 2020. Vol. 34. No 05. P. 7732–7739. DOI: 10.1609/aaai.v34i05.6276.
32. Al-Smadi M., Al-Zboon S., Jararweh Y., Juola P. Transfer learning for Arabic named entity recognition with deep neural networks // IEEE Access. 2020. Vol. 8. P. 37736–37745. DOI: 10.1109/ACCESS.2020.2973319.
33. Triggerner: Learning with entity triggers as explanations for named entity recognition / B.Y. Lin [et al.] // arXiv:2004.07493v4. 2020.
34. Cho M., Ha J., Park C., Park S. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition // Journal of Biomedical Informatics. 2020. Vol. 103. Article ID 103381. DOI: 10.1016/j.jbi.2020.103381.

35. Neural architectures for named entity recognition / G. Lample [et al.] // arXiv:1603.01360v3. 2016.

36. Huang Z., Xu W., Yu K. Bidirectional LSTM-CRF models for sequence tagging // arXiv:1508.01991v1. 2015.

37. Universal sentence encoder / D. Cer [et al.] // arXiv:1803.11175v2. 2018.

38. Multilingual universal sentence encoder for semantic retrieval / Y. Yang [et al.] // arXiv:1907.04307v1. 2019.

39. LSTM: A search space Odyssey / K. Greff [et al.] // IEEE Transactions on Neural Networks and Learning Systems. 2016. Vol. 28. No 10. P. 2222–2232. DOI: 10.1109/TNNLS.2016.2582924.

40. Song Y., Kim E., Lee G.G., Yi B.-K. POSBIOTM-NER in the shared task of BioNLP/NLPBA2004 // Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP). Geneva, Switzerland, 28–29 August 2004. P. 103–106.

Об авторах

Кузьмин Владимир Никифорович

доктор военных наук, профессор;

ведущий научный сотрудник военного института (научно-исследовательского) Военно-космической академии имени А.Ф. Можайского, 197198, г. Санкт-Петербург, ул. Ждановская, д. 13;

E-mail: vka@mil.ru

ORCID: 0000-0002-6411-4336

Менисов Артем Бакытжанович

кандидат технических наук;

докторант Военно-космической академии имени А.Ф. Можайского, 197198, г. Санкт-Петербург, ул. Ждановская, д. 13;

E-mail: vka@mil.ru

ORCID: 0000-0002-9955-2694

An approach to identifying threats of extracting confidential data from automated control systems based on internet technologies

Vladimir N. Kuzmin

E-mail: vka@mil.ru

Artem B. Menisov

E-mail: vka@mil.ru

Space Military Academy named after A.F. Mozhaysky

Address: 13, Zhdanovskaya Street, Saint Petersburg 197198, Russia

Abstract

Together with ubiquitous, global digitalization, cybercrime is growing and developing rapidly. The state considers the creation of an environment conducive to information security to be a strategic goal for the development of the information society in Russia. However, the question of how the "state of protection of the individual, society and the state from internal and external information threats" should be achieved in accordance with the "Information Security" and the "Digital Economy of Russia 2024" programs remains open. The aim of this study is to increase the efficiency whereby automated control systems identify confidential data from html-pages to reduce the risk of using this data in

the preparatory and initial stages of attacks on the infrastructure of government organizations. The article describes an approach that has been developed to identify confidential data based on the combination of several neural network technologies: a universal sentence encoder and a neural network recurrent architecture of bidirectional long-term short-term memory. The results of an assessment in comparison with modern means of natural language text processing (SpaCy) showed the merits and prospects of the practical application of the methodological approach.

Key words: information security; countering information security threats; confidential data; personal data; machine learning; deep learning; identifying the entities of natural language texts.

Citation: Kuzmin V.N., Menisov A.B. (2021) An approach to identifying threats of extracting confidential data from automated control systems based on internet technologies. *Business Informatics*, vol. 15, no 3, pp. 35–47. DOI: 10.17323/2587-814X.2021.3.35.47

References

1. *The Information Security Doctrine of the Russian Federation*. Approved by Decree of the President of the Russian Federation No 646 of 5 December 2016. Available at: http://pravo.gov.ru/proxy/ips/?docbody=&prevDoc=102161033&backlink=1&&nd=102417017 (accessed 01 February 2021) (in Russian).
2. Shaydullina V.K. (2019) Big data and personal data protection: the main theoretical and practical issues of legal regulation. *Society: Politics, Economics, Law*, no 1, pp. 51–55 (in Russian). DOI: 10.24158/pep.2019.1.8.
3. Federal Law of the Russian Federation No 152-FZ of 27 July 2006 "*About personal data*." Available at: http://pravo.gov.ru/proxy/ips/?docbody&nd=102108261 (accessed 01 February 2021) (in Russian).
4. Decree of the President of the Russian Federation No 188 of 06 March 1997 "*On approval of the list of confidential information*." Available at: http://pravo.gov.ru/proxy/ips/?docbody=&firstDoc=1&lastDoc=1&nd=102046005 (accessed 01 February 2021) (in Russian).
5. Federal Law of the Russian Federation No 79-FZ of 27 July 2004 "*About the state civil service of the Russian Federation*." Available at: http://pravo.gov.ru/proxy/ips/?docbody=&firstDoc=1&lastDoc=1&nd=102088054 (accessed 01 February 2021) (in Russian).
6. Information message "*On the development of the methodological document of the FSTEC of Russia* "*Methodology for determining information security threats in information systems*" No 240/22/1534 of 9 April 2020. Available at: https://fstec.ru/normotvorcheskaya/informatsionnye-i-analiticheskie-materialy/2071-informatsionnoe-soobshchenie-fstek-rossii-ot-9-aprelya-2020-g-n-240-22-1534 (accessed 01 February 2021) (in Russian).
7. Los V.P., Nikulchev E.V., Pushkin P.Yu., Rusakov A.M. (2020) Information-analytical system of monitoring implementation of legislation requirements by operators of personal information. *Information Security Problems. Computer Systems*, no 3, pp. 16–23 (in Russian).
8. Kozin I.S. (2018) Providing personal data protection in an information system based on user behavior analytics. *Information and Control Systems,* no 3, pp. 69–78 (in Russian). DOI: 10.15217/issn1684-8853.2018.3.69.
9. Ivichev V.A., Ignatova T.V. (2013) Technologies of identification and depersonalization of Data. *ECO*, vol. 43, no 2, pp. 168–179 (in Russian).
10. Platonov A.A., Psaryov A.A (2019) Approach to identifying program defects based on mutually complementing semantics. *Proceedings of the Mozhaisky Military Space Academy*, no 666, pp. 148–157 (in Russian).
11. Grishman R. (1994) Whither written language evaluation? Proceedings of the *Workshop on Human Language Technology. Plainsboro, New Jersey, 8–11 March 1994*, pp. 120–125.
12. Minkov E., Wang R.C., Cohen W. (2005) Extracting personal names from email: Applying named entity recognition to informal text. Proceedings of the *Human Language Technology Conference and the Conference on Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada, 6–8 October 2005*, pp. 443–450.
13. Grishman R. (1995) The NYU System for MUC-6 or Where's the Syntax? Proceedings of the *Sixth Message Understanding Conference (MUC-6). Columbia, Maryland, 6–8 November 1995*, pp. 167–175.
14. Wakao T., Gaizauskas R., Wilks Y. (1996) Evaluation of an algorithm for the recognition and classification of proper names. Proceedings of the *16th International Conference on Computational Linguistics (COLING 1996). Copenhagen, 5–9 August 1996*, vol. 1, pp. 418–423.
15. Alfred R., Leong L.C., On C.K., Anthony P. (2014) Malay named entity recognition based on rule-based approach. *International Journal of Machine Learning and Computing*, vol. 4, no 3, pp. 300–306. DOI: 10.7763/IJMLC.2014.V4.428.
16. Salleh M.S., Asmai S.A., Basiron H., Ahmad S. (2018) Named entity recognition using fuzzy c-means clustering method for Malay textual data analysis. *Journal of Telecommunication, Electronic and Computer Engineering*, vol. 10, no 2–7, pp. 121–126.
17. Zhou G.D., Su J. (2002) Named entity recognition using an HMM-based chunk tagger. Proceedings of the *40th Annual Meeting of the Association for Computational Linguistics. Philadelphia, PA, USA, 6–12 July 2002*, pp. 473–480.
18. Morwal S., Jahan N., Chopra D. (2012) Named entity recognition using hidden Markov model (HMM). *International Journal on Natural Language Computing*, vol. 1, no 4, pp. 15–23. DOI: 10.5121/ijnlc.2012.1402.
19. Morwal S., Jahan N. (2013) Named entity recognition using hidden Markov model (HMM): an experimental result on Hindi, Urdu and Marathi languages. *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 3, no 4, pp. 671–675.
20. Borthwick A. (1999) *A maximum entropy approach to named entity recognition* (PhD Thesis). New York: New York University.

21. Chieu H.L., Ng H.T. (2002) Named entity recognition: a maximum entropy approach using global information. Proceedings of the *19th International Conference on Computational Linguistics (COLING 2002). Taipei, Taiwan, 24 August – 1 September 2002*, vol. 1, pp. 1–7. DOI: 10.3115/1072228.1072253.

22. Chieu H.L., Ng H.T. (2003) Named entity recognition with a maximum entropy approach. Proceedings of the *Seventh Conference on Natural language learning at HLT-NAACL 2003. Edmonton Canada, 31 May 2003*, pp. 160–163.

23. Speck R., Ngomo A.C.N. (2014) Ensemble learning for named entity recognition. Proceedings of the *13th International Semantic Web Conference (ISWC 2014). Riva del Garda, Italy, 19–23 October 2014*, pp. 519–534.

24. Paliouras G., Karkaletsis V., Petasis G., Spyropoulos C.D. (2000) Learning decision trees for named-entity recognition and classification. Proceedings of the *ECAI Workshop on Machine Learning for Information Extraction (ECAI 2000). Berlin, 21 August 2000*, pp. 1–6.

25. Szarvas G., Farkas R., Kocsor A. (2006) A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms. Proceedings of the *9th International Conference on Discovery Science (DS 2006). Barcelona, Spain, 7–10 October 2006*, pp. 267–278. DOI: 10.1007/11893318_27.

26. Mansouri A., Affendey L. S., Mamat A. (2008) Named entity recognition approaches. *International Journal of Computer Science and Network Security*, vol. 8, no 2, pp. 339–344.

27. Ekbal A., Bandyopadhyay S. (2010) Named entity recognition using support vector machine: A language independent approach. *International Journal of Electrical, Computer, and Systems Engineering*, vol. 4, no 2, pp. 155–170.

28. Li J., Sun A., Han J., Li C. (2020) A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge and Data Engineering* (Early access). DOI: 10.1109/TKDE.2020.2981314.

29. Ma X., Hovy E. (2016) End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv:1603.01354v5.*

30. Li P.-H., Dong R.-P., Wang Y.-S., Chou J.-C., Ma W.-Y. (2017) Leveraging linguistic structures for named entity recognition with bidirectional recursive neural networks. Proceedings of the *2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 7–11 September 2017*, pp. 2664–2669.

31. Fu J., Liu P., Zhang Q. (2020) Rethinking generalization of neural models: A named entity recognition case study. Proceedings of the *34th AAAI Conference on Artificial Intelligence. New York, USA, 7–12 February 2020*, vol. 34, no 05, pp. 7732–7739. DOI: 10.1609/aaai.v34i05.6276.

32. Al-Smadi M., Al-Zboon S., Jararweh Y., Juola P. (2020) Transfer learning for Arabic named entity recognition with deep neural networks. *IEEE Access*, vol. 8, pp. 37736–37745. DOI: 10.1109/ACCESS.2020.2973319.

33. Lin B.Y., Lee D.-H., Shen M., Moreno R., Huang X., Shiralkar P., Ren X. (2020) Triggerner: Learning with entity triggers as explanations for named entity recognition *arXiv:2004.07493v4.*

34. Cho M., Ha J., Park C., Park S. (2020) Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. *Journal of Biomedical Informatics*, vol. 103, article ID 103381. DOI: 10.1016/j.jbi.2020.103381.

35. Lample G., Ballesteros M., Subramanian S., Kawakami K., Dyer C. (2016) Neural architectures for named entity recognition. *arXiv:1603.01360v3.*

36. Huang Z., Xu W., Yu K. (2015) Bidirectional LSTM-CRF models for sequence tagging. *arXiv:1508.01991v1.*

37. Cer D., Yang Y., Kong S.-Y., Hua N., Limtiaco N., John R.S., Constant N., Guajardo-Cespedes M., Yuan S., Tar C., Sung Y.-H., Strope B., Kurzweil R. (2018) Universal sentence encoder. *arXiv:1803.11175v2.*

38. Yang Y., Cer D., Ahmad A., Guo M., Law J., Constant N., Abrego G.H., Yuan S., Tar C., Sung Y.-H., Strope B., Kurzweil R. (2019) Multilingual universal sentence encoder for semantic retrieval. *arXiv:1907.04307v1.*

39. Greff K., Srivastava R.K., Koutník J., Steunebrink B.R., Schmidhuber J. (2016) LSTM: A search space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no 10, pp. 2222–2232. DOI: 10.1109/TNNLS.2016.2582924.

40. Song Y., Kim E., Lee G.G., Yi B.-K. (2004) POSBIOTM-NER in the shared task of BioNLP/NLPBA2004. Proceedings of the *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP). Geneva, Switzerland, 28–29* August *2004*, pp. 103–106.

About the authors

Vladimir N. Kuzmin

Dr. Sci. (Mil.), Professor;

Leading Researcher, Military Institute (Science and Researching), Space Military Academy named after A.F. Mozhaysky, 13, Zhdanovskaya Street, Saint Petersburg 197198, Russia;

E-mail: vka@mil.ru

ORCID: 0000-0002-6411-4336

Artem B. Menisov

Cand. Sci. (Tech.);

Doctoral Student, Space Military Academy named after A.F. Mozhaysky, 13, Zhdanovskaya Street, Saint Petersburg 197198, Russia;

E-mail: vka@mil.ru

ORCID: 0000-0002-9955-2694