УДК 004.934.1

СРАВНИТЕЛЬНЫЙ АНАЛИЗ СИСТЕМ ИЗВЛЕЧЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ ИЗ НЕСТРУКТУРИРОВАННЫХ ПУБЛИЦИСТИЧЕСКИХ ТЕКСТОВ

А.А. Навроцкий
Заведующий кафедрой информационных технологий автоматизированных систем БГУИР, кандидат физико-математических наук, доцент,

Е.В. Кривальцевич
Аспирант кафедры информационных технологий автоматизированных систем БГУИР,

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: navrotsky@bsuir.by, elena.krivaltsevich@gmail.com

А.А. Навроцкий
Окончил Белорусский государственный университет. Заведующий кафедрой информационных технологий автоматизированных систем БГУИР, кандидат физико-математических наук, доцент.

Е.В. Кривальцевич
Окончила Белорусский государственный университет. Аспирант БГУИР.

Аннотация. Рассматриваются особенности извлечения именованных сущностей из неструктурированных публицистических текстов на английском языке, находящихся в открытом доступе, посредством различных программных решений: GATE, NLTK, SpaCy, Apache OpenNLP, Stanford NER, Intellexer NER API, Thomson Reuters Open Calais API.

Ключевые слова: именованная сущность, неструктурированный текст.

Стремительный прирост информации каждый день является вызовом для существующих систем автоматизированной обработки текста. Чем больше информации, тем более точными могут быть результаты анализа, или предсказательных моделей, но, в то же время, затраты на обработку гигантских массивов информации требуют соответствующих мощностей, оптимизации скорости выполнения задач и т.д. По данным ресурса «WorldoMeters» [1] ежедневно публикуется более 5000 новых названий книг, более 4,500,000 постов в блогах, более 560,000,000 сообщений в социальной сети «Twitter», а также распространяется более 370,000,000 экземпляров газет. В большинстве случаев публикуемая информация не является закрытой для общего использования, что означает возможность ее экстрагирования и последующего статистического анализа, установления социально-экономических зависимостей, для экспериментов с предсказательными моделями и других областях с последующим использованием в процессах и алгоритмах программного обеспечения для принятия решений.

Для достижения максимально точных результатов обработки неструктурированного текста необходимы корпусы текстов высокого качества. Также для корпусов актуальна

интерпретируемость, достоверность, релевантность и точность, в том числе качественная разметка «золотого стандарта» [2].

Существуют различные программные комплексы, библиотеки, программы и программные интерфейсы приложений (API) для решения задачи извлечения именованных сущностей. Примеры:

– General Architecture for Text Engineering (GATE) – программный комплекс для интеллектуального анализа данных. Включает в себя различные компоненты, в том числе систему токенизации, частеречное тегирование, сегментацию по предложениям, извлечение именованных сущностей и анализ кореферентности. Используются алгоритмы конечных автоматов и регулярные выражения [3];

– Natural Language Toolkit (NLTK) – пакет библиотек и программ для символьной и статистической обработки естественного языка [4];

– SpaCy – библиотека для углубленной обработки текстов на естественных языках, с использованием токенизации, статистических нейронных сетей и мультиязычного извлечения именованных сущностей. Также поддерживается глубинное обучение для использования статистических моделей, созданных с помощью таких программных решений, как TensorFlow, Keras, Scikit-learn и PyTorch [5];

– Apache OpenNLP – библиотека для обработки текста на естественном языке с использованием таких методов, как определение языка, токенизация, сегментация, выделение именованных сущностей, синтаксический анализ, разбиение по ключевым словам [6];

– Stanford NER (CRFClassifier) – программа для извлечения именованных сущностей, использующая реализацию моделей последовательностей условного случайного поля (Conditional Random Field, CRF) с линейной цепью произвольного порядка. Данная реализация обеспечивает возможность использования того же кода для построения моделей последовательностей как для задачи извлечения именованных сущностей, так и для других задач посредством обучения моделей на размеченных данных [7];

– Intellexer Named Entity Recognizer API – программный интерфейс приложения (API) для извлечения именованных сущностей, в частности, названий организаций, имен собственных, географических названий, позиций (занятости), национальностей, дат, возраста, длительности, названий событий. Используется статистическая модель, основанная на скрытой модели Маркова, генерации паттернов именованных сущностей с использованием машинного обучения, а также более 500 правил, вручную заданных экспертами [8];

– Thomson Reuters Open Calais API – программный интерфейс приложения (API) для выделения именованных сущностей, в частности, имен собственных, организаций, локаций. [9].

Качество извлечения именованных сущностей вышеобозначенными программами, программными комплексами, библиотеками и программными интерфейсами приложений (далее – *программные решения*) при обработке корпуса текстов на одном естественном языке зависит от количества типов выделяемых именованных сущностей, полноты онтологий и словарей, корректности и полноты лингвистических правил, а также от качества тренировочных данных для решений, использующих машинное обучение. При обработке текстов, написанных с использованием сочетаний разных естественных языков программные решения должны быть оснащены модулями распознавания языка и всеми соответствующими, максимально полными модулями для выделения именованных сущностей.

В данном исследовании программные решения были протестированы на корпусе публицистических текстов (100 новостных статей), собранных вручную с новостного сайта «BBC» [10].

Исследовалось количество типов, выделяемых каждым решением, количество выделяемых именованных сущностей, количество именованных сущностей после проверки

экспертом (проверка экспертом включала в себя удаление некорректно выделенных элементов, корректировку переносов выделенных элементов).

В различных программных решениях используются различные подходы к выделению групп и классов именованных сущностей (Таблица 1).

Таблица 1. – Типы именованных сущностей, выделяемых различными программными решениями

№	Название программного решения	Типы именованных сущностей (как указано в документации к программному решению)	Типы именованных сущностей (описательный перевод на основе документации к программному решению)
1	ANNIE GATE	1 Person; 2 Location; 3 Organization; 4 Money; 5 Percent; 6 Date; 7 Address; 8 Identifier; 9 Unknown.	– Персонажи, люди; – Локации; – Организации; – Денежные единицы; – Проценты; – Даты; – Адреса; – Определители; не использовался в исследовании; – Именованные сущности, которые не были отнесены ни к одному из типов; не использовался в исследовании.
2	NLTK	– Person; – Location; – Organization; – Money; – Percent; – Date; – Time; – Facility; – GPE, Geo-political entities.	– Персонажи, люди; – Локации; – Организации; – Деньги; – Проценты; – Даты; – Промежутки времени, начиная от суток и менее; – Памятники культуры; – Города, штаты, провинции, страны.
3	Apache OpenNLP	– Person; – Date; – Time; – Locations; – Organizations; – Money; – Percentage.	– Персонажи, люди; – Даты; – Промежутки времен, начиная от суток и менее; – Локации; – Организации; – Денежные единицы; – Проценты.
4	Stanford NER	– Person; – Location; – Organization; – Date; – Time; – Money; – Percentage.	– Персонажи, люди; – Локации; – Организации; – Даты; – Временные промежутки, начиная от суток и менее; – Денежные единицы; – Проценты.

Продолжение *таблицы 1*

5	SpaCy	– Person; – Norp; – Fac; – Org; – GPE; – Loc; – Product; – Event; – Work_of_art; – Law; – Language; – Date; – Time; – Percent; – Money; – Quantity; – Ordinal; – Cardinal.	– Персонажи, люди; – Национальности, религиозные или политические группы; – Строения, аэропорты, дороги, мосты и т.д.; – Компании, агентства, институты и т.д.; – Страны, города, штаты; – Локации, не относящиеся к типу №5 (GPE), горные цепи, водоемы; – Объекты, устройства, пищевые продукты и т.д., исключая сервисы; – Ураганы, имеющие название, войны, сражения, спортивные события и т.д.; – Названия книг, песен и т.д.; – Названия документов, имеющих законодательную силу; – Названия языков; – Даты, периоды; – Промежутки времени, начиная от суток и менее; – Процент; – Денежные единицы; – Измерения, такие как вес и расстояние; – Порядковые числительные; – Числительные, которые не относятся ни к одному из вышеперечисленных типов.
6	Intellexer NER API	– Person; – Loc; – Loc-misc; – Org; – Event; – Nat; – Age; – Date; – Dur; – Title; – Url; – Pos.	– Персонажи, люди; – Локации; – Памятники культуры и иные географические объекты, не входящие в тип №2 (Loc); – Организации; – События; – Национальности; – Возраст; – Даты; – Промежутки времени; – Названия; – Ссылки; – Должности.

Проодолжение *таблицы 1*

7	Thomson Reuters Open Calais API	– Anniversary; – City; – Company; – Continent; – Country; – Editor; – EmailAddress; – EntertainmentAwardEvent; – Facility; – FaxNumber; – Holiday; – IndustryTerm; – Journalist; – MarketIndex; – MedicalCondition; – MedicalTreatment; – Movie; – MusicAlbum; – MusicGroup; – NaturalFeature; – OperatingSystem; – Organization; – Person; – PharmaceuticalDrug; – PhoneNumber; – PoliticalEvent; – Position; – Product; – ProgrammingLanguage; – ProvinceOrState; – PublishedMedium; – RadioProgram; – RadioStation; – Region; – SportsEvent; – SportsGame; – SportsLeague; – Technology; – TVShow; – TVStation; – URL.	– Годовщины; – Города; – Компании; – Континенты; – Страны; – Редакторы; – Адреса электронной почты; – Награждения; – Памятники культуры; – Номера факса; – Праздники; – Термины, специфические для индустрии; – Журналисты; – Индексы рынка акций и ценных бумаг; – Названия медицинских состояний; – Названия способов лечения; – Кинофильмы; – Музыкальные альбомы; – Музыкальные группы; – Реки, озера, моря, океаны, горы, названия географических регионов; – Операционные системы; – Организации; – Люди и персонажи; – Названия лекарств; – Номера телефонов; – Политические события; – Должности; – Продукты; – Языки программирования; – Провинции, штаты; – Названия газет, журналов и иных периодических изданий; – Названия радиопрограмм; – Названия радиостанций; – Регионы; – Спортивные события; – Спортивные игры; – Спортивные лиги; – Названия технологий; – ТВ-шоу и телепрограммы; – ТВ-станции и каналы; – Веб-ссылки.

Стоит отметить, что корректность определения типа именованной сущности в данном исследовании не учитывается. Также в силу различности определений и трактовок типов именованных сущностей в рамках разных программных решений, было принято решение не

учитывать данные различия. Это означает, что два решения, один из которых выделяет, к примеру, 5 типов именованных сущностей, а другой – 7 типов, считаются эквивалентными.

Данный тип сравнения может приводить к тому, что программные решения, не выделяющие определенный тип именованных сущностей будут проигрывать при общем сравнении с другими решениями. Поэтому был выбрано сравнение по количеству некорректно выделенных сущностей в рамках одного и того же решения. Некорректно выделенные сущности осложняют анализ полученных результатов и затрудняют работу с последующими автоматизированными решениями (калькуляция статистики, визуализация и т.д.). Поэтому важно иметь максимально чистые выходные данные для последующих процедур.

Анализ эффективности программных решений в рамках одного и того же типа именованных сущностей будет рассмотрен в последующих исследованиях.

В Таблице 2 приведено сравнение по проценту некорректно выделенных именованных сущностей при обработке 100 публицистических текстов на английском языке.

Таблица 2. – Сравнение программных по результатам обработки 100 неструктурированных публицистических текстов на английском языке

Название решения	**Всего выделено именн.сущ.тей**	**Всего именн.сущ.-тей после ручной проверки**	**% некорректно выделенных именн.сущ.-тей**	**Точность (Precision)**
ANNIE GATE	5299	5117	3,43%	0.97
NLTK	10283	5760	43,99%	0.56
SpaCy	7445	7388	0,77%	0.99
Apache OpenNLP	3143	2017	35,83%	0.64
Stanford NER	12259	8965	26,87%	0.73
Intellexer NER API	3712	3639	1,97%	0.98
Thomson Reuters Open Calais API	2469	2378	3,69%	0.97

Согласно полученным данным, одним из наиболее точно работающих решений является SpaCy. Данное программное решение поддерживает возможность не только использования заранее натренированной модели [11], но и обучения модели на данных исследователя [12]. Программное решение SpaCy NER может найти свое применение в методах и алгоритмах естественно-языковых систем, в частности, диалоговых систем и экспертных систем.

Список литературы

[1] WorldoMeter Statistics [Электронный ресурс] – Режим доступа: http://www.worldometers.info/

[2] Assessing the Quality of Unstructured Data: An Initial Overview [Электронный ресурс] – Режим доступа: https://pdfs.semanticscholar.org/a7d3/1b09498a16201f7044eb77ff14d27c1c559b.pdf

[3] Named Entity Recognition with ANNIE [Электронный ресурс] – Режим доступа: http://services.gate.ac.uk/annie/

[4] Natural Language Toolkit [Электронный ресурс] – Режим доступа: https://www.nltk.org/

[5] Industrial-Strength Natural Language Processing [Электронный ресурс] – Режим доступа: https://spacy.io/

[6] Apache OpenNLP [Электронный ресурс] – Режим доступа: https://opennlp.apache.org/

[7] Stanford Named Entity Recognizer (NER) [Электронный ресурс] – Режим доступа: https://nlp.stanford.edu/software/CRF-NER.html

[8] Intellexer Named Entity Recognizer [Электронный ресурс] – Режим доступа: https://www.intellexer.com/ner.html

[9] Refinitiv Intelligent Tagging [Электронный ресурс] – Режим доступа: http://www.opencalais.com/opencalais-api/

[10] BBC News [Электронный ресурс] – Режим доступа: https://www.bbc.com/news

[11] Simple and Accurate Dependency Parsing Using Bidirectional LSTM Feature Representations [Электронный ресурс] – Режим доступа: https://www.semanticscholar.org/paper/Simple-and-Accurate-Dependency-Parsing-Using-LSTM-Kiperwasser-Goldberg/eec3a236ecd185712ce65fb336141f8656eea13d

[12] Training spaCy's Statistical Models [Электронный ресурс] – Режим доступа: https://spacy.io/usage/training

COMPARATIVE ANALYSIS OF SYSTEMS FOR EXTRACTING NAMED ENTITIES FROM UNSTRUCTURED PUBLICISTIC TEXTS

A. NAVROTSKY
Head of the Department of Information Technologies in Automated Systems in BSUIR, Candidate of Physics and Mathematics, Associate Professor

E. KRIVALTSEVICH
Postgraduate student of Information Technologies in Automated Systems in BSUIR

Belarusian State University of informatics and radioelectronics, Republic of Belarus
E-mail: navrotsky@bsuir.by, elena.krivaltsevich@gmail.com

Abstract. The features of extracting named entities from unstructured publicistic English texts in the public domain using various software solutions: GATE, NLTK, SpaCy, Apache OpenNLP, Stanford NER, Intellexer NER API, Thomson Reuters Open Calais API are considered.

Keywords: named entity, unstructured text.