МЕТОДЫ ИДЕНТИФИКАЦИИ ИМЕНОВАННЫХ СУЩНОСТЕЙ В ЗАДАЧАХ ОБРАБОТКИ ПОТОКА НАУЧНЫХ НОВОСТЕЙ

METHODS FOR IDENTIFICATION OF NAMED ENTITIES IN PROCESSING THE STREAM OF SCIENTIFIC NEWS

Костюк Даниил Максимович *– младший научный сотрудник, Государственная публичная научно-техническая библиотека СО РАН (Российская Федерация), kostuk@gpntbsib.ru,*

Широков Никита Константинович *– младший научный сотрудник, Государственная публичная научно-техническая библиотека СО РАН (Российская Федерация), e-mail: shirokov@gpntbsib.ru*

Kostuk Daniil Maksimovich *– junior researcher, The State Public Science and Technology Library, kostuk@gpntbsib.ru*

Shirokov Nikita Konstantinovich *– junior researcher, The State Public Science and Technology Library, e-mail: shirokov@gpntbsib.ru*

Аннотация: *В рамках одного из проектов фундаментальных научных исследований ГПНТБ СО РАН планируется расширение научно-информационного ресурса «Новости сибирской науки» на российскую науку в целом и добавление аналитического блока, обеспечивающего отдельным ученым и научным организациям представление об этом аспекте значимости их работы. Для решения этой задачи необходимо выделять из новостного потока именованные сущности (названия статей, упоминания авторов, организаций и т.д.). Это позволит улучшить качество отбора новостных сообщений, построить необходимую аналитику. Именованные сущности позволяют также решать проблемы сопоставления перепечаток, классификации новостных сообщений и автоматизации работы.*

Abstract: *Within the framework of one of the fundamental scientific research projects of the State Public Scientific Technical Library of the SB RAS, it is planned to expand the scientific and information resource "News of Siberian Science" to Russian science as a whole and add an analytical block that provides individual scientists and scientific organizations with an idea of this aspect of the significance of their work. To solve this problem, it is necessary to select named entities (titles of articles, mentions of authors, organizations, etc.) from the news stream. This will allow both to improve the quality of the selection of news messages and to build the necessary analytics. Named entities can also address reprint matching, newsletter classification, and work automation.*

Ключевые слова: *извлечение именованных сущностей, обработка естественного языка, анализ данных, наукометрия.*

Keywords: *named entity recognition, natural language processing data analysis, scientometrics.*

Альтернативные подходы к измерению значимости научных результатов прежде всего в широких кругах общества по сравнению с метриками, характеризующими внимание профессионального сообщества, основанными на цитировании, получили широкое распространение. Ключевым элементом этих альтметрик является анализ упоминаний в новостных сообщениях средств массовой информации. Однако инструменты, агрегирующие необходимые данные и предоставляющие сводную информацию по ним, опираются преимущественно на англоязычные ресурсы. В рамках текущей политики Российской Федерации аналитика научных организаций, персон и локаций, полученная из СМИ является важным критерием оценки их результативности и влияния.

Ресурс «Новости сибирской науки» был создан в 2015 году для представления сообщений из печатных и электронных СМИ и на данный момент содержит более 45 000 новостных сообщений. Из-за планируемого расширения научно-информационного ресурса «Новости сибирской науки» на российскую науку в целом, возникла необходимость в автоматизации обработки данных, для увеличения объёма новостных сводок за день, уменьшения количества перепечаток и упрощения работы операторов. Одним из направлений данных работ является предоставление информации об именованных сущностях, встречаемых в тексте, таких как персоны, организации и локации. Это позволит оператору не только не тратить время на сопоставление с базой данных, но использовать полученную информацию для анализа упоминаний полученных сущностей в потоке новостей. Поэтому одной из целей стала разработка алгоритма выделения именованных сущностей в потоке научных новостей.

Для решения поставленной цели необходимо было решить следующие задачи:

1. Провести анализ существующих алгоритмов идентификации именованных сущностей.
2. Реализовать механизм идентификации именованных сущностей в текстах новостных сообщений.
3. Интегрировать разработанное решение в административный интерфейс ресурса «Пульс.Науки».

Задача NER (Named Entity Recognition) состоит в поиске именованных сущностей и их последующей классификации на персоны, организации локации и т.д.

Одними из методов поиска именованных сущностей являются методы, основанные на разборе грамматик языков (Natasha) и статистических моделях (DeepPavlov).

Парсеры грамматик языка основываются на предварительно составленных наборе правил и для увеличения полноты выделяемых именованных сущностей, этими правилами необходимо описать большое количество языковых ситуаций, что бывает очень затруднительно. Однако они обладают хорошим быстродействием и простотой в исправлении ситуаций-ошибок.

В статистических моделях для решения используются классические модели машинного обучения или глубокое обучение. И хотя они показывают наиболее лучшие и стабильные результаты [1], но требуют намного больше времени в реализации, объёмную выборку для обучения и имеют большую вычислительную сложность. Детальное сравнение подобных моделей приведено в статье [2].

Для выделения именованных сущностей наш выбор был сделан в пользу библиотеки Natasha – молодого проекта для задач обработки естественного русского языка. Одним из главных преимуществ этой библиотеки является то, что она создавалась для анализа текстов новостных сообщений, используя в разы меньше памяти, при этом не уступая в качестве обработки. Данные свойства алгоритма позволяют использовать его для обработки большого потока новостных сообщений в реальном времени, что является важной составляющей будущего ресурса.

Скорость и качество библиотеки Razdel, которая занимается делением русскоязычного текста на токены и предложения, сопоставимы или выше, чем у других русскоязычных открытых решений (Рис. 1).

Библиотека Slovnet, которая занимается обучением современных моделей для русскоязычного NLP, достаточно близко подошло к качеству DeepPavlov BERT NER (Рис. 2), размер модели получился меньше в 75 раз, потребление памяти в 30 раз меньше, скорость обработки новостных статей в секунду в 2 раза больше, при этом разница F1 меры составляет 0.01 для выделенных сущностей организации, персон и локаций.

Решения для сегментации на токены	Ошибки на 1000 токенов	Время обработки, секунды
Regexp-baseline	19	**0.5**
SpaCy	17	5.4
NLTK	130	3.1
MyStem	19	4.5
Moses	**11**	**1.9**
SegTok	12	2.1
SpaCy Russian Tokenizer	**8**	46.4
RuTokenizer	15	**1.0**
Razdel	**7**	2.6

Рис. 1. Сравнение работы сегментации на популярных открытых датасетах: SynTagRus, OpenCorpora, GICRYA

	Natasha, Slovnet NER	DeepPavlov BERT NER
PER/LOC/ORG F1 по токенам, среднее по Collection5, factRuEval-2016, BSNLP-2019, Gareev	0.97/0.91/0.85	0.98/0.92/0.86
Размер модели	27МБ	2ГБ
Потребление памяти	205МБ	6ГБ (GPU)
Производительность, новостных статей в секунду (1 статья ≈ 1КБ)	25 на CPU (Core i5)	13 на GPU (RTX 2080 Ti), 1 на CPU
Время инициализации, секунд	1	35
Библиотека поддерживает	Python 3.5+, PyPy3	Python 3.6+
Зависимости	NumPy	TensorFlow

Рис. 2. Сравнение работы Slovnet NER и SOTA DeepPavlov BERT NER

В качестве данных для работы были взяты новости, представленные на ресурсе «Новости сибирской науки». Так как они имеют формат HTML, была произведена их очистка от спецсимволов и HTML-тегов, после чего исправлены опечатки и орфографические ошибки с помощью YandexSpeller. В связи с тем, что нам важен контекст выделенной именной сущности в предложении, очистка текстов от стоп-слов нам не требуется. Лемматизация именованных сущностей проводится после их выделения. Также были составлены базы данных персон, организаций и локаций, используемых в новостных сообщениях.

Была составлена тестовая выборка, состоящая из 5 700 новостей институтов, с размеченными полями именованных сущностей, для проверки работы выделения ORG (организаций) алгоритмом. В качестве оценки точности работы алгоритма мы будем использовать F1 оценку, т.е. гармоническое среднее точности и полноты [3]. Применение стандартного алгоритма Natasha без последующей обработки результатов показало очень низкие результаты (Рис. 3). В процессе анализа результатов работы, были выявлены следующие ошибки:

- Неправильно проведённая лемматизация слов.
- Разбиение названия одной организации на две.
- Различные формы написания часто встречаемых и общепринятых аббревиатур и сокращений.
- Недостаточно вариативный словарь из базы данных организаций.

	F1 Мера	Precision	Recall
Natasha	0.51	0.67	0.41
Natasha + Доп. обработка	0.79	0.82	0.76

Рис. 3. Оценки точности Natahsa с дополнительной обработкой и без на примере ORG

Проведя работу по написанию инструкций и правил, учитывающих данные ошибки, F1 оценку удалось увеличить на 0.28 (Рис. 3).

СПИСОК ЛИТЕРАТУРЫ

1. Neural Architectures for Named Entity Recognition [Electronic resource] / Guillaume Lample [at al.] // Arxiv.org. – Mode of access: https://arxiv.org/pdf/1603.01360.pdf. – Date of access: 12.04.2021.

2. Можарова, В. А. Двухэтапный подход к извлечению именованных сущностей / В. А. Можарова, Н. В. Лукашевич // Пятнадцатая национальная конференция по искусственному интеллекту с международным участием, 3–7 окт. 2016 г., г. Смоленск, Россия : КИИ-2016 : тр. конф. : [в 3 т.]. – Смоленк, 2016. – Т. 2. – С. 81–88.

3. Mansouri, A. Named Entity Recognition Approaches [Electronic resouce] / Alireza Mansouri, Lilly Suriani Affendey, Ali Mamat // Intern. J. of Computer Science and Network Security. – 2008. – Vol. 8, iss. 2. – Mode of access: http://paper.ijcsns.org/07_book/200802/20080246.pdf. – Date of access: 12.12.2009.