Е. В. Кривальцевич

АВТОМАТИЗИРОВАННОЕ ИЗВЛЕЧЕНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ ИЗ НЕСТРУКТУРИРОВАННЫХ ПУБЛИЦИСТИЧЕСКИХ ТЕКСТОВ

Рассматривается извлечение именованных сущностей из неструктурированных публицистических текстов на английском языке посредством различных программных решений.

I. Введение

Широко известно, что в настоящее время идёт активный ежедневный прирост информационных ресурсов. Одним из применений данной информации может быть анализ и последующее использование в системах принятия решений.

II. Сравнительный анализ существующих систем

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

1) General Architecture for Text Engineering (GATE) [1];
2) Natural Language Toolkit (NLTK) [2];
3) SpaCy [3];
4) Apache OpenNLP [4];
5) Stanford NER (CRFClassifier) [5];
6) Intellexer NER API [6];
7) Thomson Reuters Open Calais API [7].

Таблица 1 – Сравнение эффективности программных решений для выделения именованных сущностей: I – Название решения, II – Всего выделено именн.сущ.-тей, II – Всего именн.сущ.-тей после ручной проверки, IV – Кол-во некорректно выделенные именн.сущ.-тей

I	II	III	IV
NLTK	10283	5760	43,99%
SpaCy	7445	7388	0,77%
Apache OpenNLP	3143	2017	35,83%
Stanford NER	12259	8965	26,87%
Intellexer NER API	3712	3639	1,97%
Thomson Reuters API	2469	2378	3,69%

В данном исследовании программные решения были протестированы на корпусе публицистических текстов (100 новостных статей), собранных вручную с новостного сайта «BBC» [8].

В таблице (Таблица 1) приведено сравнение по проценту некорректно выделенных именованных сущностей при обработке 100 публицистических текстов на английском языке.

III. Выводы

Согласно полученным данным, одним из наиболее точно работающих решений является Intellexer NER API. Демонстрационным приложением, использующим Intellexer NER API является приложение News Monitoring [9].

1. GATE's ANNIE System [Electronic resource] – Mode of access : http://services.gate.ac.uk/annie/. – Date of access : 08.04.2019.
2. Natural Language Toolkit [Electronic resource] – Mode of access : https://www.nltk.org/. – Date of access : 08.04.2019.
3. spaCy [Electronic resource] – Mode of access : https://spacy.io/. – Date of access : 08.04.2019
4. Apache OpenNLP [Electronic resource] – Mode of access : https://opennlp.apache.org/. – Date of access : 08.04.2019.
5. Stanford Named Entity Recognizer (NER) [Electronic resource] – Mode of access : https://nlp.stanford.edu/software/CRF-NER.html/. – Date of access : 08.04.2019.
6. Intellexer Named Entity Recognizer API [Electronic resource] – Mode of access : https://www.intellexer.com/ner.html. – Date of access : 08.04.2019.
7. Thomson Reuters Open Calais™ API [Electronic resource] – Mode of access : http://www.opencalais.com/opencalais-api/. – Date of access : 08.04.2019.
8. BBC News Services [Electronic resource] – Mode of access : https://www.bbc.com/news. – Date of access : 08.04.2019.
9. Intellexer News Monitoring System[Electronic resource] – Mode of access : http://newsmap.intellexer.com. – Date of access : 08.04.2019.

Кривальцевич Елена Викторовна, кафедра информационных технологий автоматизированных систем, БГУИР, elena.krivaltsevich@gmail.com.

Научный руководитель: Навроцкий Анатолий Александрович, заведующий кафедрой информационных технологий автоматизированных систем, кандидат физико-математических наук, доцент, navrotsky@bsuir.by.