Н. В. Зыгмантович

РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ В ТЕКСТЕ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

С самого начала эпохи компьютеров разработчики стараются научить их понимать обычные языки. В течение тысяч лет люди что-то писали, и было бы здорово поручить машинам чтение и разбор всех этих данных.

Конечно, мечта о понимании естественного языка пока еще далека от полноценной реализации. Однако некоторые результаты есть уже сейчас: машину можно научить обнаруживать нужные объекты в тексте на естественном языке, находить между ними связи и представлять необходимые данные в формализованном виде для дальнейшей обработки.

Именованные сущности – это слова или словосочетания, которые обозначают предметы или явления определенной категории. В соответствии с поставленными задачами в текстах могут выделяться имена и фамилии людей, названия организаций, населенных пунктов и других географических объектов, даты, денежные единицы и т.д. Задача извлечения именованных сущностей является подзадачей задачи извлечения информации.

В процессе извлечения именованных сущностей можно выделить два этапа:

- распознавание именованной сущности;
- категоризация именованной сущности.

На первом этапе происходит обнаружение слова или цепочки слов, которые образуют сущность. Каждое слово представлено токеном: *«Пирамиды Гиза»* – это цепочка из токенов, которые репрезентируют одну сущность. Специалисты пользуются тегированием для того, чтобы продемонстрировать, где сущности начинаются и заканчиваются.

На втором этапе создаются категории сущностей.

Для того, чтобы понять, какие сущности важны и релевантные, а какие нет и как их соотносить с различными категориями, нужны обучающие данные (training data). Чем более релевантными являются обучающие данные, тем более точными и верными будут результаты.

Задача извлечения именованных сущностей была сформулирована более 20 лет назад. За это время был разработан не один подход к ею решению. Изначально задачу решали, основываясь на составленных вручную правилах, однако для этого от исследователя требовалось иметь высокую квалификацию в области грамматики языка, при этом количество используемых языков было ограниченным, а списки рассматриваемых слов постоянно нуждались в обновлении.

Позже стали применяться методы машинного обучения и другие, каждый из которых имеет свои достоинства и недостатки.

Факторы, влияющие на системы извлечения именованных сущностей:

- языковой фактор;
- жанры и предметные области текстов.

При разработке систем извлечения именованных сущностей для конкретного языка учитываются его особенности. Например, в английском языке тексты читаются слева направо, а к именованным сущностям часто относятся имена собственные, которые пишутся с заглавной буквы. Однако нельзя применять те же методы к таким языкам, как фарси, иврит, где чтение ведется справа налево, а в начале имен собственных нет заглавных букв.

Тексты относятся к разным стилям речи (публицистический, научный, разговорный), посвящены разным областям (например, политика, медицина, наука, экономика, спорт). Особенности стилей и предметных областей учитываются в системах извлечения именованных сущностей, специализированных на конкретных типах текстов. Примерами таких типов являются электронные письма, научные статьи, новостные заметки, записи телефонных разговоров и другие. Экспериментально установлено, что система, хорошо работающая с текстами определенного типа, показывает худшие результаты при обработке текстов другого типа.

К основным типам именованных сущностей обычно относят PERSON, ORGANIZATION, LOCATION. Также рассматриваются категории Timex, которая включает типы «время», «дата», и Numex, которая включает типы

«денежное выражение», «процентное выражение». В зависимости от предметных областей и приложений системы извлечения именованных сущностей могут добавляться новые типы именованных сущностей.

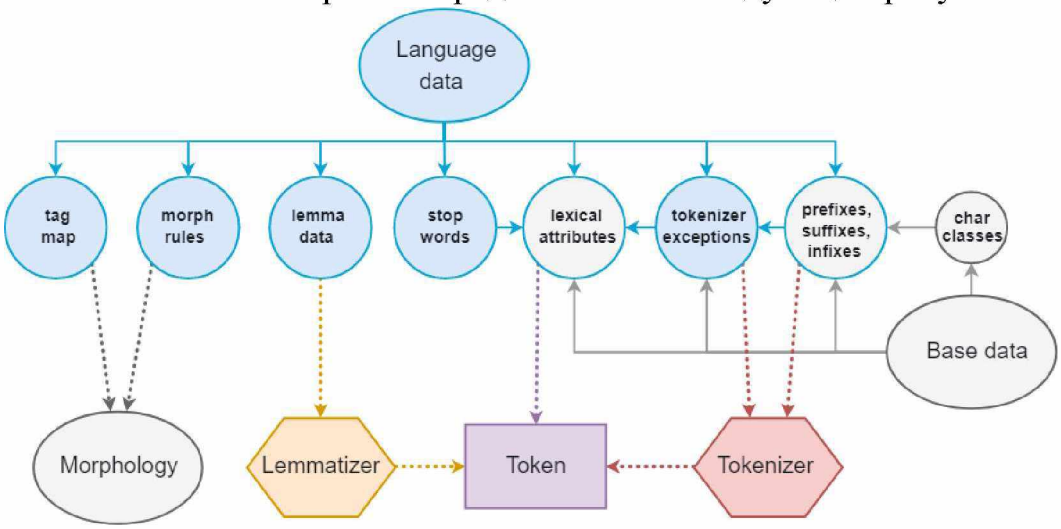
Проведем компьютерный эксперимент, напишем небольшой код на языке программирования Python и на практике извлечем именованные сущности из текста на английском языке.

Существуют библиотеки языка Python для выполнения различных задач обработки естественного языка – **textacy**, **spaCy**, в которых уже реализованы такие шаги, как токенизация, теги частей речи, разбор зависимостей и т. д. Именно эти шаги необходимо предварительно выполнить для успешного решения нашей задачи. Попробуем воспользоваться этими библиотеками.

В языке программирования Python установим эти библиотеки:

```
# Установка spaCy , textacy
!pip install -U spacy
!pip install -U textacy
```

Загрузим в память ПК из интернета готовую языковую модель "*en_core_web_lg*" на английском языке, в которой уже реализованы этапы решения задач NLP. Схематично ее работа представлена на следующем рисунке:



Чтобы загрузить эту модель в память нашего ПК, составим следующий код:

```
import spacy.cli
spacy.cli.download("en_core_web_lg")
import en_core_web_lg
nlp = en_core_web_lg.load()
```

«Забросим» свой текст в эту готовую модель и получим нужные нам результаты.

Возьмем текст на английском языке и попробуем извлечь из него именованные сущности. Для чего напишем следующий код:

```
text = """Jerusalem (CNN) If Israel and Hamas agree on anything at the mo
ment, it is probably that Gaza is as close to another war as it has been since the en
```

d of the last one four years ago. And the next one, predicts Yahya Sinwar, Hamas' Gaza leader, in an interview published in the Israeli daily Yedioth Ahronoth, will li kely be the most severe. "It cannot end like the third [war], which ended like the se cond [war], which ended like the first [war]," Sinwar told reporter Francesca Bor ri, who conducted the interview. The Israelis, he said, "will have to reoccupy Gaza. " But the Hamas leader also said that a new war is not in anybody's interests. This is the first full interview given by Sinwar to non-Arabic
media since he became the leader of Hamas in Gaza
in February 2017. Read More"""

```
doc = nlp(text)#здесь теперь содержится обработанная версия текста
# распечатаем все обнаруженные именованные сущности
for entity in doc.ents:
    print(f"{entity.text} ({entity.label_})")
```

Команды, которые начинаются со знака #, описывают пояснения. Запустив этот код, мы получим результат:

Jerusalem (GPE)
CNN (ORG)
Israel (GPE)
Hamas (ORG)
Gaza (GPE)
the end of the last one four years ago (DATE)
Yahya Sinwar (PERSON)
Hamas (ORG)
Gaza (GPE)
Israeli (NORP)
daily (DATE)
Yedioth Ahronoth (PERSON)
third (ORDINAL)
second (ORDINAL)
first (ORDINAL)
Sinwar (ORG)
Francesca Borri (PERSON)
Israelis (NORP)
Gaza (GPE)
Hamas (ORG)
first (ORDINAL)
Sinwar (ORG)
non-Arabic (NORP)
Hamas (ORG)
Gaza (GPE)
February 2017 (DATE)

Вот так легко и просто мы извлекли структурированную информацию из неструктурированного теста, получив именованные сущности данного текста.