Введение

Данная задача относится к задаче извлечения информации (IE) в области обработки естественного языка (NLP). Её можно разбить на две подзадачи:

* Распознавание именованных элементов (выделить основных действующих лиц и организации)
* Автоматическое реферирование (определить суть события)

Будем решать её с помощью методов машинного обучения (библиотека STRANZA, NLTK и spaCy).

Основная часть

# **Алгоритм выделения активных субъектов**

Выделение людей и организаций в тексте реализуем с помощью библиотеки STRANZA. Данная библиотека поддерживает русский язык и способна решать задачу распознавания именованных элементов (NER).

Алгоритм делится на следующие этапы:

* Создаем Pipeline
* Подаем текст на вход
* Для каждого предложения в тексте находим все сущности
* Если тип сущности ORG или PER (Организация или человек), то добавляем эту сущность в массив
* Удаляем все повторяющиеся сущности (set)
* Выводим результат

Ниже приведен код

def stanza\_nlp\_ru(text):

  nlp = stanza.Pipeline(lang='ru', processors='tokenize,ner')

  doc = nlp(text)

  ans = []

  for sent in doc.sentences:

    for ent in sent.ents:

      if ent.type in ['ORG', 'PER']:

        ans.append(ent.text)

        print(f'Entity: {ent.text}\tType: {ent.type}')

  return set(ans)

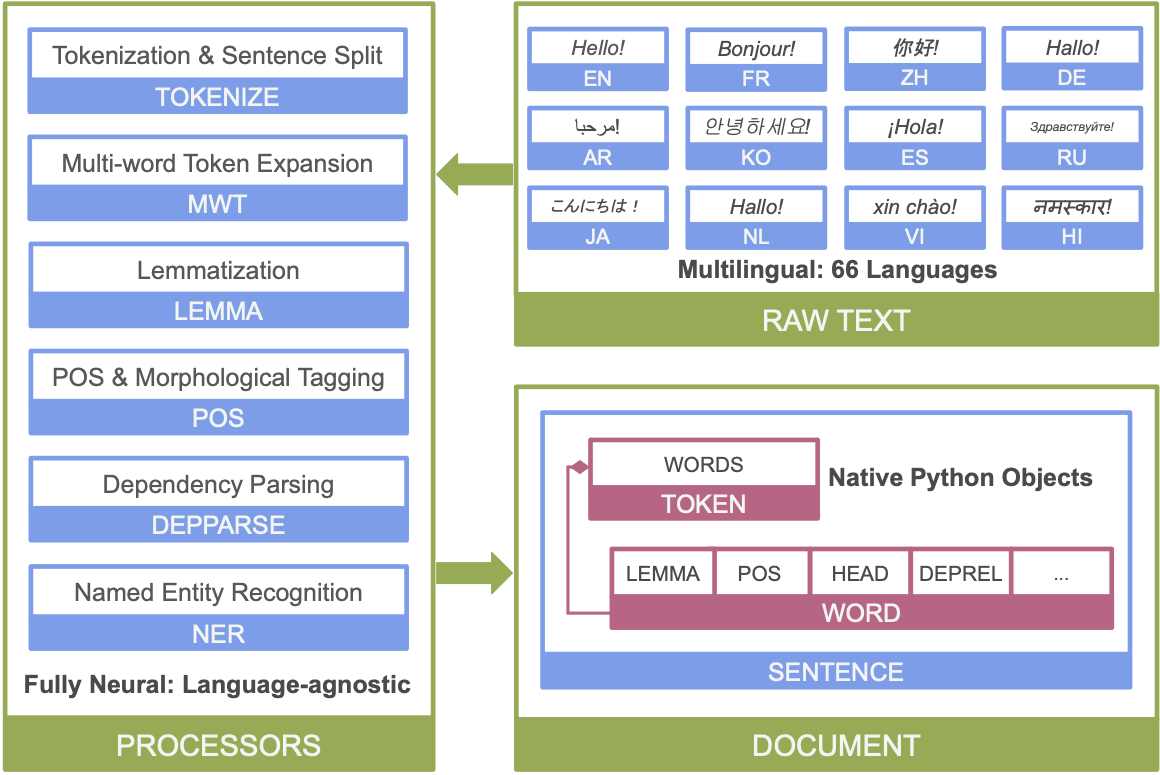


Рис. 1. Stranza pipeline

# **Алгоритм выделения сути событий**

Уффф

# **Выбор инструментов**

Реализация будет разработана на языке Python с использованием библиотек Stranza, NLTK и spaCy. Данный выбор обусловлен простотой использования и выполнением требуемых функций.

# **Программная реализация**

Для скачивания данных запустим следующий код:

import csv

with open('/content/6229a20db2ede.csv', newline='') as f:

    reader = csv.reader(f)

    data = list(reader)

Для скачивания соответствующих библиотек:

!pip install stanza

import stanza

stanza.download('ru')

Метод для выделения активных субъектов:

def stanza\_nlp\_ru(text):

  nlp = stanza.Pipeline(lang='ru', processors='tokenize,ner')

  doc = nlp(text)

  ans = []

  for sent in doc.sentences:

    for ent in sent.ents:

      if ent.type in ['ORG', 'PER']:

        ans.append(ent.text)

        print(f'Entity: {ent.text}\tType: {ent.type}')

  return set(ans)

# **Демонстрация**

Ссылка на github: https://github.com/BaronFonMonc/1

for i in range(1,10):

  print(data[i])

  russian\_text = data[i][1]

  print("Основные действующие лица: ", stanza\_nlp\_ru(russian\_text))

|  |  |
| --- | --- |
| **Текст** | **Активные субъекты** |
| Российская корпорация "Иркут" и египетский холдинг KATO Investment на авиасалоне в Дубае подписали ряд соглашений о развитии сотрудничества по проекту новейшего российского пассажирского самолета МС-21 . "В рамках программы сотрудничества заключено соглашение о закупке для авиакомпании Cairo Aviation (дочернее предприятие KATO Investment ) 6 самолетов МС-21. Соглашение также предусматривает опцион на 4 самолета МС-21" , — сообщается в понедельник на сайте российской компании. Кроме того, в настоящее время стороны обсуждают создание регионального центра по ремонту и обслуживанию МС-21 в районе международного аэропорта Аль-Аламейн, расположенного в 184 километрах от Каира. Египетская сторона планирует бесплатно предоставить площадь, примыкающую к воздушной гавани, для размещения производственных сооружений. Отметим, что в настоящее время на Иркутском авиазаводе ведется сборка первых летных прототипов МС-21. Первый полет запланирован на весну-лето будущего года, серийное производство — на 2017 год. | Cairo Aviation,Иркутском авиазаводе,Иркут,KATO Investment |
| На реализацию потребуется 13 лет, отметил гендиректор "Вертолетов России" Андрей Богинский . © Фото: пресс-служба президента РФ Россия и Китай заключили контракт на создание тяжелого вертолета, соглашение было подписано 25 июня. Об этом доложил президенту РФ Владимиру Путину генеральный директор холдинга "Вертолеты России" Андрей Богинский на встрече в Кремле. "В 2016 году в ходе Вашего визита в Пекин было подписано межправсоглашение. С 2008 года шли интенсивные переговоры, и 25 июня текущего года мы подписали контракт", - сказал он. По словам гендиректора, российская сторона будет заниматься разработкой и созданием таких агрегатов, как трансмиссия, рулевой винт и противообледенительная система, а также накачкой ресурса для данной машины. Совместный проект двух стран рассчитан на 13 лет. Также, по словам Богинского , холдинг с 2021 по 2025 годы планирует произвести свыше 1100 вертолетов. Межправительственное соглашение о создании перспективного тяжелого вертолета AHL (Advanced Heavy Lift) было подписано компанией "Вертолеты России" и китайской компанией Avicopter (входит в корпорацию Aviation Industry Corporation of China, AVIC) в Пекине в 2016 году. | Avicopter,Владимиру Путину,Aviation Industry Corporation of China,Вертолеты России,Андрей Богинский,Вертолетов России,AVIC,Богинского,AHL |
| В Воронежской области дан старт первому в России производству сельхозмашин французской компании KUHN . Торжественная церемония пуска нового завода состоялась в Рамонском районе. По поручению губернатора Александра Гусева в открытии принял участие заместитель председателя правительства Воронежской области Виктор Логинов . Он передал поздравления главы региона, который отметил важность сотрудничества с KUHN Group в рамках приоритетного стратегического проекта "Новая индустриализация региона". Открывшийся завод уже устанавливает кооперационные связи с машиностроительными предприятиями Воронежской области, осуществляет контакты с местными вузами – в плане стажировки и дальнейшего трудоустройства студентов. В дальнейшем это сотрудничество будет расширяться. Производственная платформа введенного в строй дочернего предприятия ООО "Кун Восток" состоит из трех линий мощностью до 500 машин в год. В рамках строительства первой очереди также сданы в эксплуатацию склады готовой продукции (площадью 7 тыс. кв. метров) и запасных частей (на 5 тыс. палетомест). На территории расположен профессиональный учебный центр, шоу-рум с экспозицией и испытательный полигон площадью 10 га для демонстрации техники в работе. На воронежском предприятии наладят сборку наиболее востребованных у отечественных аграриев машин – почвообрабатывающих агрегатов, зерновых и пропашных сеялок, механизмов для внесения удобрений, опрыскивания и другой техники с использованием российских комплектующих. По словам генерального директора ООО "Кун Восток" Андрея Манзюка , перед заводом поставлена задача не просто выпускать в России доступную технику под известным европейским брендом, но и совершенствовать механизмы ее послепродажной поддержки, в том числе, с использованием потенциала дилерских центров. Общий объем инвестиций в проект оценивается в 3 млрд рублей. Развитие завода "Кун Восток" будет проходить в три этапа. Уже в рамках второй очереди проекта предполагается удвоение производственной площади и склада запасных частей. Третья очередь будет включать создание цехов металлообработки, литья, сварки и покраски. | ООО "Кун Восток,Александра Гусева,Виктор Логинов,Андрея Манзюка,KUHN Group,KUHN,Кун Восток |

# **Выводы**

В целом данный алгоритм неплохо решает задачу выделения активных субъектов, но нужно обработать результаты, чтобы привести их в именительный падеж тем самым исключив ситуации, когда выводятся “Вертолетов России” и “Вертолеты России”.