Санкт-Петербургский государственный университет

Кафедра системного программирования

Автоматическая подстановка параметров текстовых макросов в корректной форме

Карими Хурматулла, Группа 22.M07-мм

Научный руководитель: к.ф.-м.н. Д.В.Лутцив, доцент Кафедра системного программирования

Санкт-Петербург
2022

Введение

- Веб сайты, использующие язык разметки
  1. Википедия
  2. Github
  3. ReStructuredText
- Для пнглийского языка подстановка простая
- Для русского подстановка требует менять форму словосочетания

Цель работы

Разработать и внедрить систему автоматической замены терминов в тексте в требуемой форме на русском языке.

- Изучить инструменты анализа тектов на естственных языках, позволяющие анализировать предложения.
- Выбрать модели машинного обучения, которые позволят, обучившись на корректных текстах выбирать нужные формы для подстановки фрагментов в текст.
- Провести эксперименты с этими моделями, выбрать наиболее подходящую.
- Реализовать прототип инструмента, выполняющего макроподстановку словосочетаний в корректном падеже и числе.

Инструменты

Инструменты, которые мы рассмотрели для анализа текстов:

- NLTK (Natural language toolkit).
- Pymorphy2
- LemmInflect

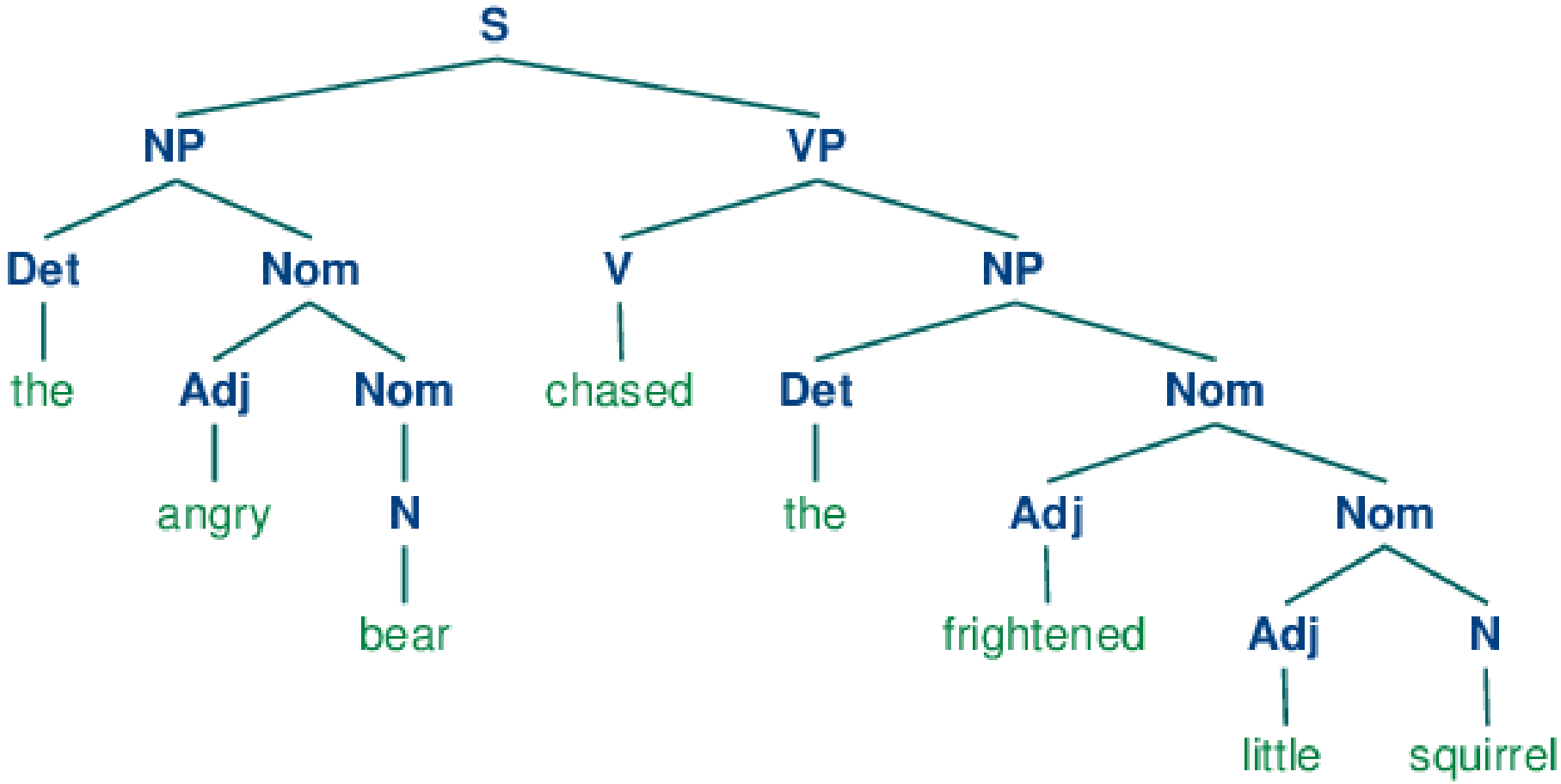
NLTK (Инструментарий естественного языка)[1]

Это платформа, которая создает программы на Python, работающие с данными на человеческом языке. NLTK содержит библиотеки классификации текста, кластеризации и обработки текста для обозначения, синтаксического анализа, классификации и создания частей речи из наших данных. NLTK слишком велик, чтобы объяснять, но мы сосредоточимся на очень специфической области нашей работы:

- Рекурсивная структура для структуры грамматики.
- Рекурсивный-нисходящий-синтаксический анализ для разбора структурированной грамматики

[1] https://www.nltk.org/

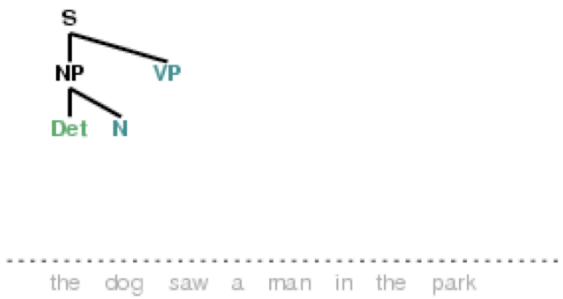
Рекурсивная структура



Рекурсивный-нисходящий-синтаксический анализ

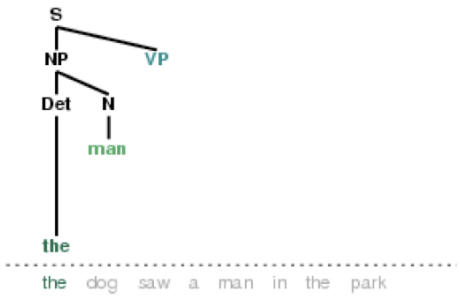
1. Initial stage

2. Second production

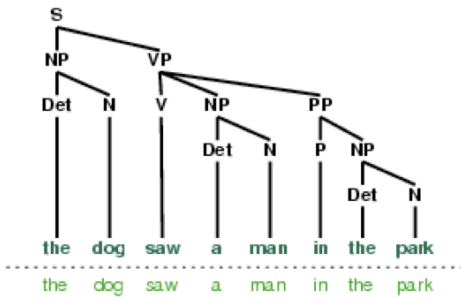


3. Matching *the*

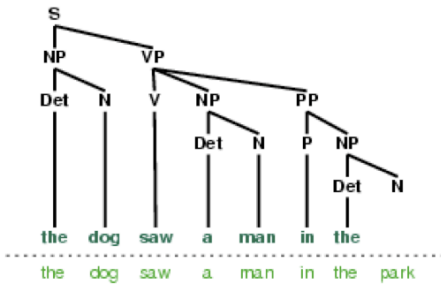
4. Cannot match *man*



5. Completed parse



6. Backtracking



Pymorphy2

Pymorphy2 это морфологический анализатор, который анализирует русские тексты с помощью словаря opencorpora. Алгоритм Pymorphy выполняет морфологическую обработку на основе грамматической характеристики типа (слова, лемматизация), но если слово не существует в словаре, поэтому предиктор в pymorphy2 объединит два алгоритма:

1. По префиксу
2. В конце слов

В то же время, когда мы анализируем слова, мы столкнемся с несколькими состояниями слов и оценками, затем, выбрав состояние слова с наивысшим баллом Мы можем исправить предложение

LemmInflect

LemmInflect использует словарный подход для лемматизации английских слов и преобразования их в формы, заданные предоставленным пользователем тегом Universal Dependencies или Penn Treebank. Библиотека работает со словами, не входящими в словарный запас (OOV), применяя методы нейронных сетей для классификации словоформ и выбора соответствующих правил морфинга.

Следующий!

- Выбрать модели машинного обучения, которые позволят, обучившись на корректных текстах выбирать нужные формы для подстановки фрагментов в текст.
- Провести эксперименты с этими моделями, выбрать наиболее подходящую.
- Реализовать прототип инструмента, выполняющего макроподстановку словосочетаний в корректном падеже и числе.

справки

- Bird, Steven. "NLTK: the natural language toolkit."Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions. 2006.
- https://lemminflect.readthedocs.io/en/latest/
- https://pymorphy2.readthedocs.io/en/stable/