Обработка естественного языка

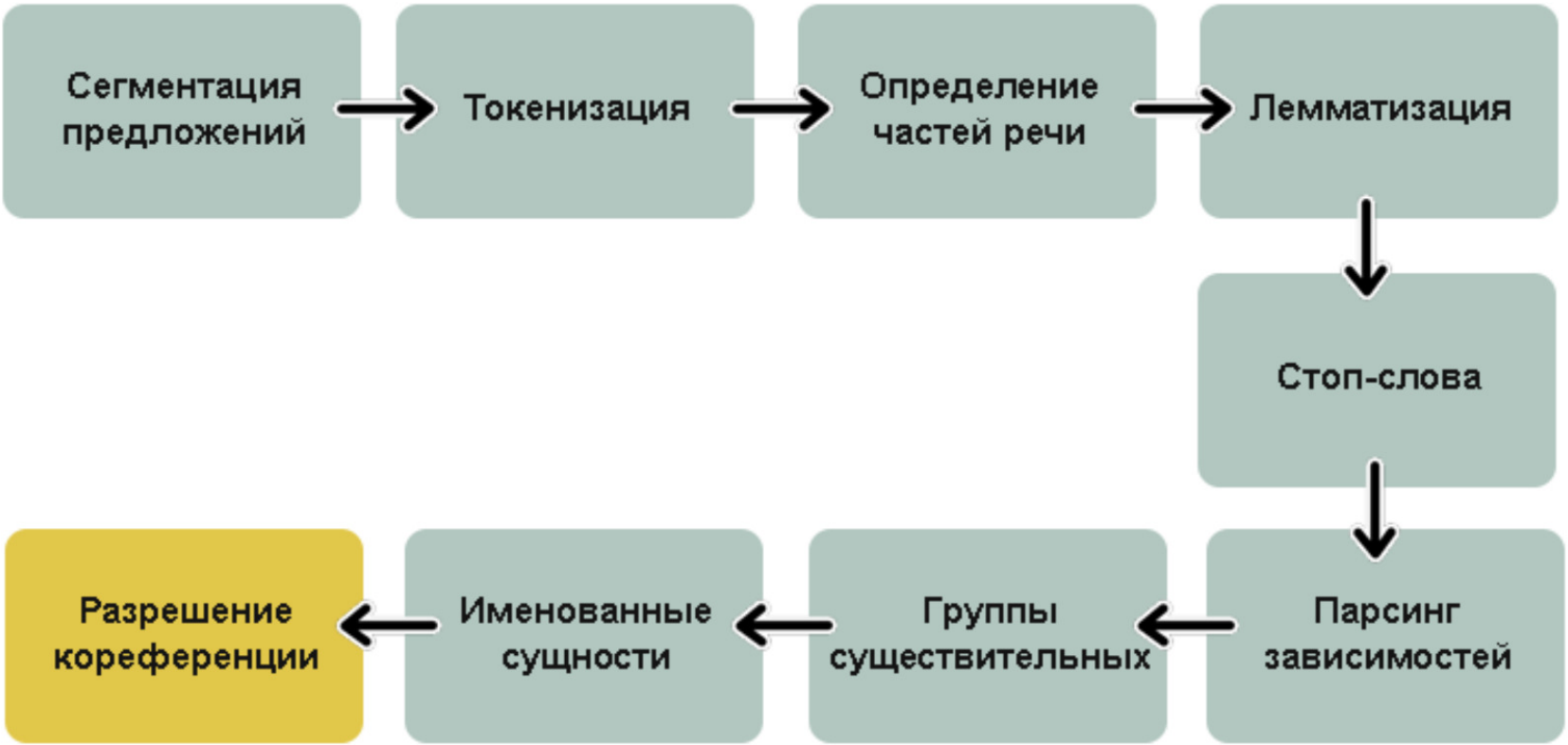
октябрь 2018

Содержание

Решаемые задачи

- Формирование ответов на вопросы (Question Answering)
- Анализ эмоциональной окраски высказываний (Sentiment Analysis)
- Нахождение текста, соответствующего изображению (Image to Text Mappings)
- Машинный перевод (Machine Translation)
- Распознавание речи (Speech Recognition)
- Извлечение сущностей (Name Entity Recognition)
- И многие другие задачи...

Этапы решения задачи обработки естественного языка

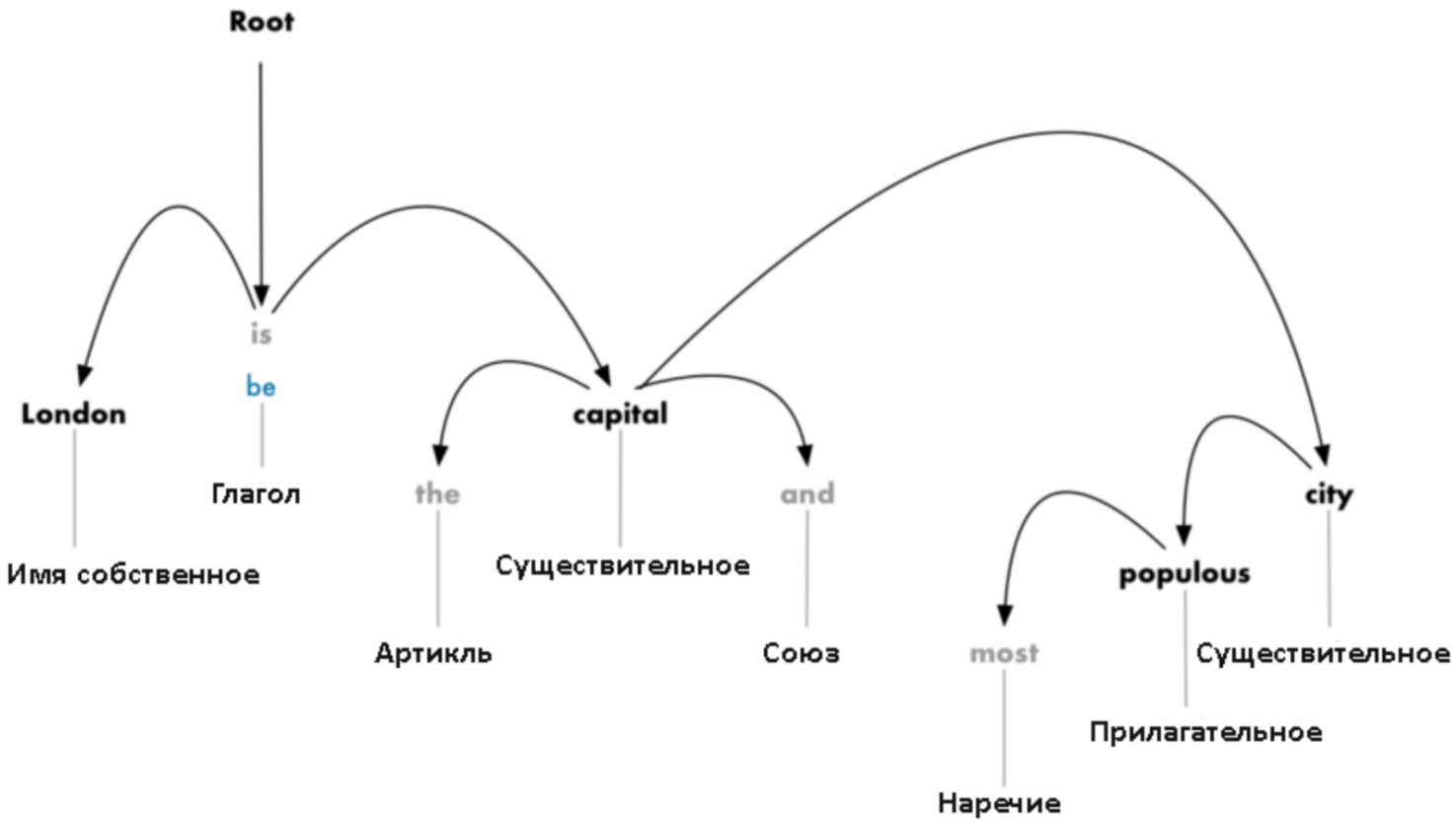


Этапы решения задачи обработки естественного языка

- **Выделение предложений**
- **Токенизация**
  Как правило, под токенизацией понимают выделение слов, но поскольку поскольку могут иметь важное значение, в некоторых моделях они тоже являются токенами.
- **Определение частей речи**
- **Лемматизация**
  Лемматизацией называется нахождением основной формы (леммы) каждого слова в предложении.
- **Определение стоп-слов**
  Как правило, можно просто отбросить слова, встречающиеся слишком часто.

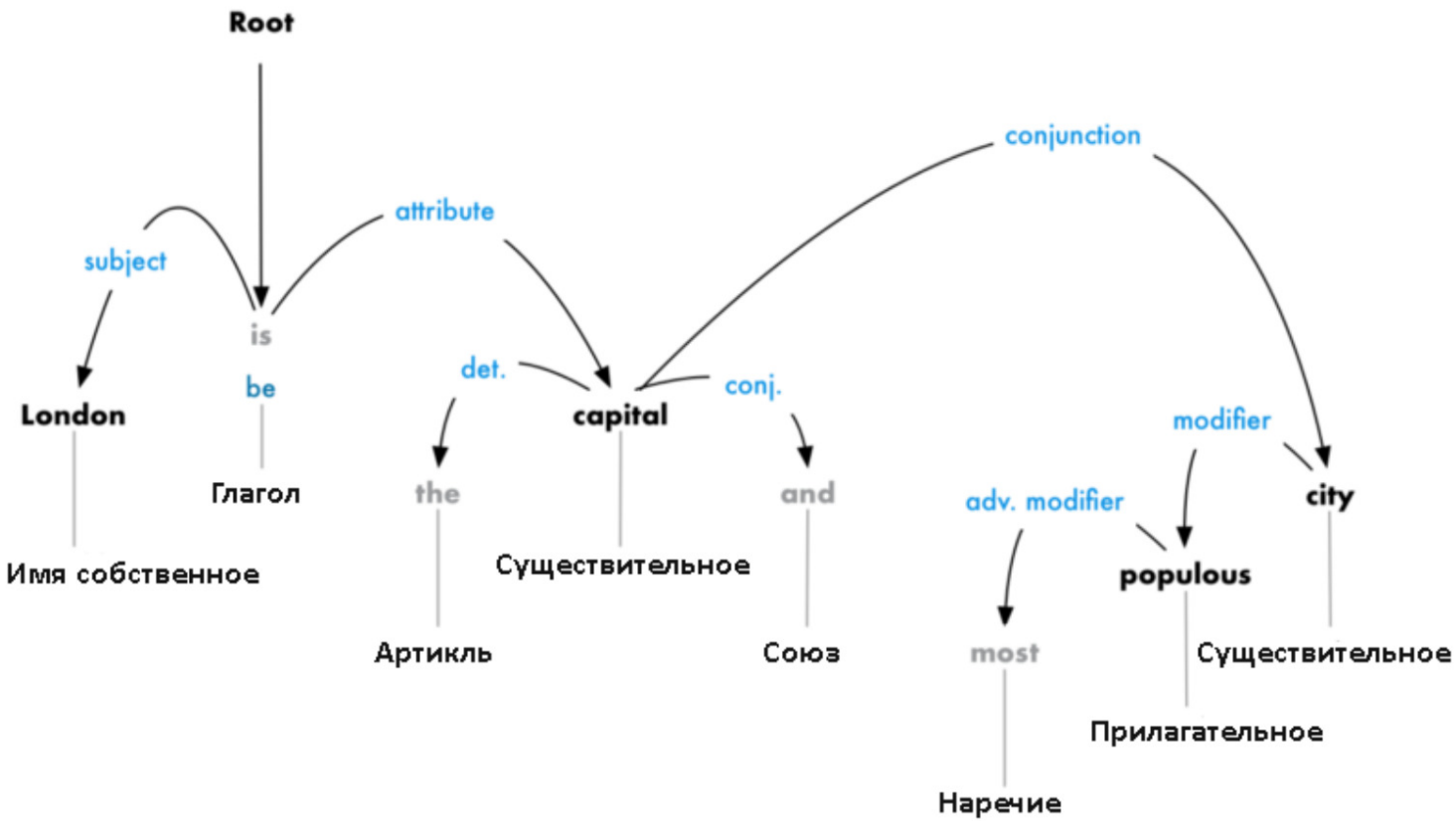
Этапы решения задачи обработки естественного языка

- Парсинг зависимостей



Этапы решения задачи обработки естественного языка

- Парсинг зависимостей



Этапы решения задачи обработки естественного языка

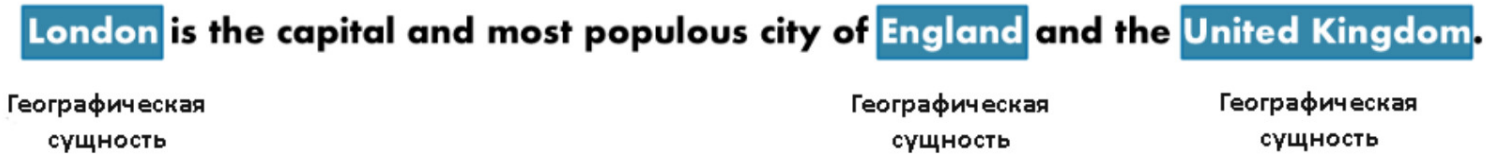
- **Поиск групп существительных**

Вместо

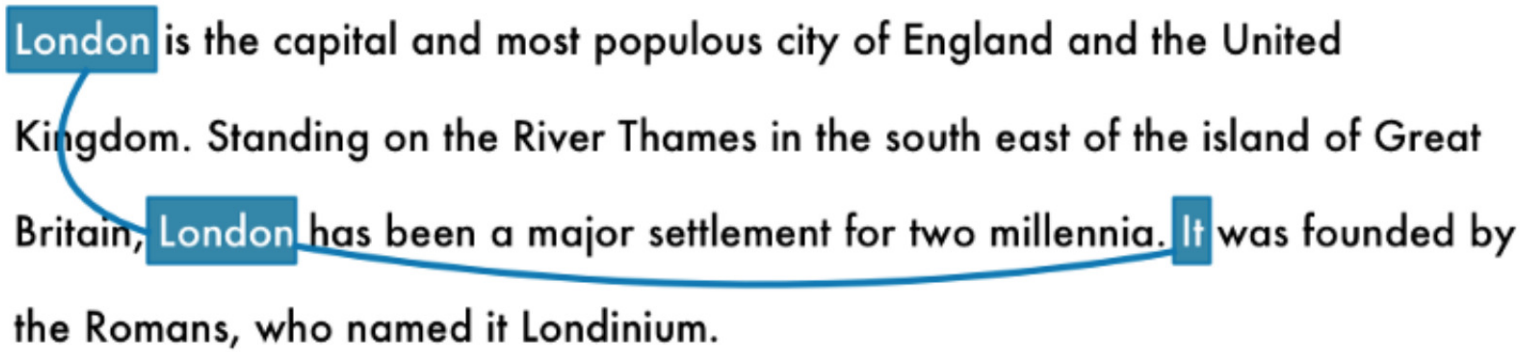
получим

Этапы решения задачи обработки естественного языка

- Распознавание именованных сущностей



- Разрешение кореференции



Еще несколько слов о предварительной обработке

- Приведенный выше путь является примерным. Часто некоторые его этапы опускаются или меняются местами в зависимости от решаемой задачи.
- Чистка текста, поиск и исправление опечаток.
- Сложность алгоритмов зависит от языка.
- При работе со стоп-словами могут возникнуть проблемы. Пример: известная группа 80-х называется «The The!»
- Технические подробности того, как именно мы обращаемся со словарями, могут быть очень существенными.

word2vec

Обозначим $w_1, w_2, \dots w_N$ обучающую последовательность слов, а c — размер контекста.

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{-c\leqslant j\leqslant c, j\neq 0}\log p(w_{n+j}|w_j) \to max$$

Обозначим v_w векторное представление слова w.

$$p(w_o|w_i) = \frac{exp(v_{w_o}^T v_{w_i})}{\sum_{k=1}^{W} exp(v_{w_k}^T v_{w_i})}$$

где W — длина словаря.

Особенности word2vec

- Максимизируется логарифмическая вероятность встречаемости слов контекста для данного центрального слова.
- Векторы преобразуя методом стохастического градиентного спуска.
- Возникают линейные отношения между разными векторами слов.

$$u_{\text{рубашка}} - u_{\text{одежда}} \approx u_{\text{стул}} - u_{\text{мебель}}$$

$$u_{\text{король}} - u_{\text{мужчина}} \approx u_{\text{королева}} - u_{\text{женщина}}$$

Языковое моделирование

Вероятностная модель:

Какова вероятность следующего слова?

$$p(house|this\ is\ the) = ?$$

Какова вероятность всей последовательности?

$$p(this\ is\ the\ house) = ?$$

- Исправление опечаток
- Распознавание рукописного текста
- Автоматические ответы
- ...

Языковое моделирование

Дана последовательность слов $s = \{w_1, w_2, \ldots w_k\}$.

Правило условной вероятности:

$$p(s) = p(w_1)p(w_2|w_1)\ldots p(w_k|w_1\ldots w_{k-1})$$

Предположение Маркова:

$$p(w_i|w_1\ldots w_{i-1}) = p(w_i|w_{i-n+1}\ldots w_{i-1})$$

Пример: биграммная модель ($n = 2$)

$$p(s) = p(w_1)p(w_2|w_1)\ldots p(w_k|w_{k-1})$$

Анализ тональности

Три типа задач (по возрастанию сложности):

- Полярная тональность (positive / negative / neutral)
- Ранжированная тональность («звёздочки» от 1 до N)
- Более сложные типы

Ключевые моменты:

- «не» лучше приклеивать к впередистоящему слову, инвертируя его тональность
- очень важно выделять смайлы и экспрессивную пунктуацию (регулярные выражения)
- обучение и тестирование нужно производить на схожих данных

Сложности

Помимо чисто технических проблем, возникают также более сложные семантические:

- Отзывы могут иметь ясный смысл, но при этом не содержать позитивных или негативных слов:

  Это фильм заставляет прочувствовать всю гамму эмоций от «А» до «Я».

- Отзыв может содержать позитивные слова, но на самом деле выражает ожидание:

  Это фильм должен был быть супер крутым. Но не был.

Подходы к решению

- Правила (точно, трудозатратно)

  Пример: если сказуемое в группе положительных глаголов и нет отрицаний то positive

- Словари (просто, зависит от предметной области)
- ML: обучение с учителем (классификация) (точно при достаточной обучающей выборке, требуется данные для обучения)
- ML: обучение без учителя (просто, не требуются данные для обучения, нужен словарь, низкая точность)

Выводы

- Задачи чаще всего предполагают много этапов решения
- Некоторые этапы требуют существенного знания языка
- Сложность алгоритмов зависит от языка
- Многие модели построены на анализе условных вероятностей