Задачи NLP: разбор

Катя Герасименко

Зимняя олимпиадная школа МФТИ

06.01.2019

В программе

- Sentiment analysis
- NER

За кадром в этот раз остаются:

- машинный перевод
- распознавание речи
- синтаксический и семантический парсинг
- чат-боты

Анализ тональности
Sentiment analysis

Что надо делать

Определить, какое высказывание: негативное, позитивное или нейтральное

Нужно, например, компаниям для оценки своей деятельности (собрать посты в соцсетях, посмотреть их тональность, сделать выводы)

Сложности

- Отрицание
- Ирония / сарказм
- несколько мнений / несколько объектов в одном предложении
- жаргонизмы, новые слова, которые не встречались в обучающем корпусе
- смайлики

Усложнение

- Предсказывать тональность конкретных слов и учитывать тональность контекста
- Сначала выделять предложения, где есть мнения, а затем классифицировать только их
- Сперва делить вход на «аспекты» - выделить конкретные сущности, для которых во входе может быть мнение. Пример: для ресторана аспектами могут быть качество еды, расположение, обслуживание и так далее.

Как можно делать

Rule-based approach: каким-нибудь образом собираем списки слов и выражений для текстов разной тональности, проверяем по итоговым спискам тональность + правила

Преимущества:

- сначала быстро и несложно
- интерпретируемо

Недостатки:

- потом долго и сложно (правила имеют свойство разрастаться в огромного неуправляемого монстра)
- правила никогда не могут охватить всё и это исправляется только новыми правилами (+ конфликты)

Как можно делать

Supervised machine learning: классификация. Если обучаем на шкале и хотим шкалу, то можно попробовать регрессию

Unsupervised machine learning: кластеризация

Признаки:

- слова, словесные n-граммы, символьные n-граммы (борьба с опечатками и богатой морфологией)
- плотные эмбеддинги – предобученные / доучить / выучивать вместе с основной задачей

Модели:

- Традиционные классификаторы
- Сейчас – чаще CNN, DCNN, LSTM

Извлечение сущностей
Entity recognition

Что надо делать

Из текста вытащить сущности самого разного рода – имена, организации, деньги, даты, названия товаров, локации, адреса, etc.

Нужно для всего – автозаполнение документов и анкет, подбор релевантных документов (фильтрация), первый этап при извлечении более сложных вещей из текста (связей, например).

Как можно делать

Простые сущности: правила

Мало данных: правила, предобученные модели

Много размеченных данных: традиционное МО

Очень много размеченных данных: нейросеть

Правила

- регулярочки))) номера телефонов, например или другие хорошо и понятно структурированные последовательности
- грамматики, составленные из правил, и парсеры для них для эффективной обработки текста
- списки (список стран, список городов, список сокращений, список имен…) + списки «ключевых слов» («рубль», «федерация», «улица» и тд)
- Для русского – парсер Yargy (можно писать свои правила), библиотека Natasha (распознавалки имен, локаций, денег, адресов, дат…)

Правила

Жирный плюс: не нужны размеченные данные, не нужно ничего обучать (только себя, потому что надо искать и придумывать разные случаи, чтобы их обрабатывать)

Жирный минус: правила быстро разрастаются и со временем все больше напоминают костыли, но чуть что в сторону – уже не справляются

Предобученные модели

Есть обученные на больших корпусах модели, чтобы просто применять их к своим данным

Плюс: не нужно ничего обучать, а умное машинное обучение есть :)

Минус:

- не работает, если данные сильно отличаются от обучающего корпуса
- черный ящик, сложно навесить фиксы
- если нейронка – начинается х а о с

Машинное обучение

Задача NER обычно формулируется как sequence labeling – нужно разметить последовательность токенов. Разметка более-менее универсальная (BIO или BILOU):

word	POS	tag
Eddy	N	B-PER
Bonte	N	I-PER
is	V	O
woordvoerder	N	O
van	Prep	O
diezelfde	Pron	O
Hogeschool	N	B-ORG
.	Punc	O

Машинное обучение

Традиционное МО
Можно делать это через обычную классификацию, можно использовать Conditional Random Fields – специальную модель для тэггирования последовательностей.

Фичи:
- сами слова (+ их эмбеддинги)
- частеречная разметка
- слова слева, слова справа
- информация о синтаксических единицах
- капитализация

Машинное обучение

Нейросети
Sequence labeling -> RNN, LSTM и другие модные слова, а на деле – нейросети, которые заточены под последовательности.