

СТАТИСТИЧЕСКИЕ ЯЗЫКОВЫЕ МОДЕЛИ

Маша Шеянова, masha.shejanova@gmail.com

August 6, 2018

НИУ ВШЭ

INTRO

ЧТО МЫ НАЗЫВАЕМ СТАТИСТИЧЕСКОЙ ЯЗЫКОВОЙ МОДЕЛЬЮ?

Определение из Википедии:

A **statistical language model** is a probability distribution over sequences of words.

По-русски:

Статистическая языковая модель — это распределение вероятностей по последовательностям слов.

ЧТО УМЕЕТ ЯЗЫКОВАЯ МОДЕЛЬ?

- Оценивать вероятность той или иной последовательности слов в языке

$$P(W) = P(w_1, \dots, w_n)$$

- Ранжировать вероятности последовательностей
- Предсказывать наиболее вероятное следующее слово при условии уже известного ряда слов

$$P(w_n | w_1, \dots, w_{n-1})$$

Статистической языковой моделью называется такая модель, которая умеет делать хотя бы один из пунктов.

Оценка вероятности последовательности: **цепное правило**:

$$P(X_1, \dots, X_n) = P(X_1)P(X_2|X_1)\dots P(X_n|X_1, \dots, X_{n-1}))$$

Вероятность следующего слова:

$$P(X_n|X_1, \dots, X_{n-1})) = \frac{P(X_1, \dots, X_n)}{P(X_1, \dots, X_{n-1}))}$$

ЗАЧЕМ ЭТО МОЖЕТ БЫТЬ ПОЛЕЗНО?

Языковые модели принождаются в огромном количестве задач:

- спеллчекинг
- автодополнение
- распознавание речи
- распознавание символов (Optical Character Recognition, OCR)
- машинный перевод
- реферирование текста
- порождение текста

ПОДХОДЫ

Мы пользуемся **марковским предположением**:

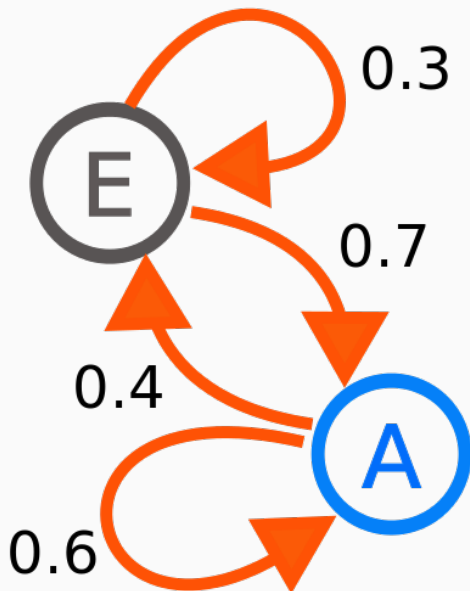
текущее состояние зависит лишь от конечного числа
предыдущих состояний

Иными словами:

$$P(w_i | w_1 \dots w_{i-1}) \approx P(w_i | w_{i-n} + 1 \dots w_{i-1})$$

Последовательность случайных событий с конечным или счётным числом исходов, характеризующаяся тем свойством, что при фиксированном настоящем будущее независимо от прошлого.

Марковская цепь — частный случай **взвешенного конечного автомата**.



Нейронные модели показывают лучшие результаты.

Бывают пословные и посимвольные.

Пословные используют word embeddings. Ключевые слова:

- RNN
- LSTM
- seq2seq

ПРИЛОЖЕНИЯ

С распознавания речи всё началось.

Что нужно для распознавания речи?

- акустическая модель (представление о фонетике языка)
- лексическая модель (о том, какими могут быть слова)
- языковая модель (вероятности последовательностей слов)

Яндекс.Рефераты (<https://yandex.ru/referats/>):

Точка перегиба оправдывает экзистенциальный принцип восприятия, открывая новые горизонты.

Плазменное образование восстанавливает элементарный платежный документ, даже с учетом публичного характера данных правоотношений.

Закон, основываясь на парадоксальном совмещении исключающих друг друга принципов характерности и поэтичности, предоставляет абстрактный голос персонажа.

ПОРОЖДЕНИЕ ТЕКСТА. ВО-ПЕРВЫХ, ЭТО ВЕСЕЛО!

Ветхий Алгоритм (https://twitter.com/alg_testament):



Ветхий Алгоритм @alg_testament · Jul 24

эта процедура должна работать не так, отец мой, отец мой, не бесчести имени Бога твоего.

Translate Tweet



1



33



94



Ветхий Алгоритм @alg_testament · Jul 22

Сигналы, генерируемые терминалом, возникают, когда пользователь входит в дом Господень

Translate Tweet



39



103



Ветхий Алгоритм @alg_testament · Jul 22

И поставлю Себе священника верного; он будет обладать правами суперпользователя.

Translate Tweet



1



82



217



ПОРОЖДЕНИЕ ТЕКСТА. ЧТО ЕЩЁ?

Чатботы!

Например, когда Алиса понимает, что пользователь хочет поговорить, она включает болталку — порождение текста.

Это может быть не только развлечение, но и психологическая помощь человеку.

Чтобы перевести последовательность слов правильно, нам мало знать самый вероятный перевод.

Важно знать, насколько вообще вероятна подобранная последовательность **в целевом языке**.

СПАСИБО ЗА ВНИМАНИЕ!
Вопросы?