Языковое моделирование
Лекция 2

Решение задачи языкового моделирования с помощью N-граммных моделей

План занятия

1. Что такое N-граммы
2. Алгоритм обучения N-граммных языковых моделей
3. N-граммные модели на практике
4. Плюсы и минусы N-граммных языковых моделей

С помощью чего можем решить эту задачу?

N-грамма — подпоследовательность из N подряд идущих слов.

С помощью чего можем решить эту задачу?

N-грамма — подпоследовательность из N подряд идущих слов.

Студенты на паре открыли книги

Униграммы: студенты, на, паре, открыли, книги

С помощью чего можем решить эту задачу?

N-грамма — подпоследовательность из N подряд идущих слов.

Студенты на паре открыли книги

Униграммы: студенты, на, паре, открыли, книги

Биграммы: студенты на, на паре, паре открыли, открыли книги

С помощью чего можем решить эту задачу?

N-грамма — подпоследовательность из N подряд идущих слов.

Студенты на паре открыли книги

Униграммы: студенты, на, паре, открыли, книги

Биграммы: студенты на, на паре, паре открыли, открыли книги

Триграммы: студенты на паре, на паре открыли, паре открыли книги

N-граммная модель

Алгоритм:

1. Выбрать значение параметра N.
2. Разбить большой текст на N-граммы.
3. Для каждой N-граммы подсчитать частоту встречаемости.
4. Сохранить N-граммные частоты.

Пример расчета вероятности

Предположим, что в нашем большом тексте следующие последовательности встретились:

студенты открыли свои — 100 раз

студенты открыли свои книги — 20 раз

студенты открыли свои ноутбуки — 80 раз

Статистики выглядят так:

$$P(\text{книги}|\text{студенты открыли свои}) = \frac{20}{100} = 0.2$$

$$P(\text{ноутбуки}|\text{студенты открыли свои}) = \frac{80}{100} = 0.8$$

N-граммные модели на практике

1. Вставить дополнительные служебные токены <s> в начало каждой последовательности в необходимом количестве, чтобы оценить вероятность первого слова

N-граммные модели на практике

1. Вставить дополнительные служебные токены <s> в начало каждой последовательности в необходимом количестве, чтобы оценить вероятность первого слова
2. Добавить <UNK> токен для слов не из словаря (Out-of-vocabulary)

   Мы вчера пошли в картинг -> <s><s> Мы вчера пошли в <UNK> </s>

N-граммные модели на практике

1. Вставить дополнительные служебные токены <s> в начало каждой последовательности в необходимом количестве, чтобы оценить вероятность первого слова
2. Добавить <UNK> токен для слов не из словаря (Out-of-vocabulary)

   Мы вчера пошли в картинг -> <s><s> Мы вчера пошли в <UNK> </s>

3. Генерировать за счет сэмплирования из вероятностного распределения на каждом шаге

N-граммные модели на практике

1. Вставить дополнительные служебные токены <s> в начало каждой последовательности в необходимом количестве, чтобы оценить вероятность первого слова
2. Добавить <UNK> токен для слов не из словаря (Out-of-vocabulary)

   Мы вчера пошли в картинг -> <s><s> Мы вчера пошли в <UNK> </s>

3. Генерировать за счет сэмплирования из вероятностного распределения на каждом шаге
4. Добавление сглаживания для избегания нулевых вероятностей для некоторых N-грамм, например, сглаживание Лапласа:

$$P_{\text{Laplace}}(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)+1}{\sum_w (C(w_{n-1}w)+1)} = \frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V}$$

Плюсы и минусы N-граммных моделей

Плюсы:

1. Очень быстрые.

   Работают за O(1)
2. Просты в обучении

Плюсы и минусы N-граммных моделей

Плюсы:

1. Очень быстрые.

   Работают за O(1)
2. Просты в обучении

Минусы:

1. Очень чувствительны к тренировочным данным
2. Занимают много памяти
3. Имеют очень короткий контекст
4. Проблема несуществующих N-грамм

Итоги занятия

1. Узнали, что такое N-граммы и как с их помощью построить статистические языковые модели
2. Познакомились с практическими аспектами обучения и использования языковых моделей
3. Узнали, что N-граммные модели очень быстрые, однако дает не лучшее качество и занимают много памяти