академия больших данных

mail.ru group

Настоящий блок лекций подготовлен при поддержке «ЦРТ | Группа компаний»

ЦРТ ГРУППА КОМПАНИЙ

Обработка речевых сигналов

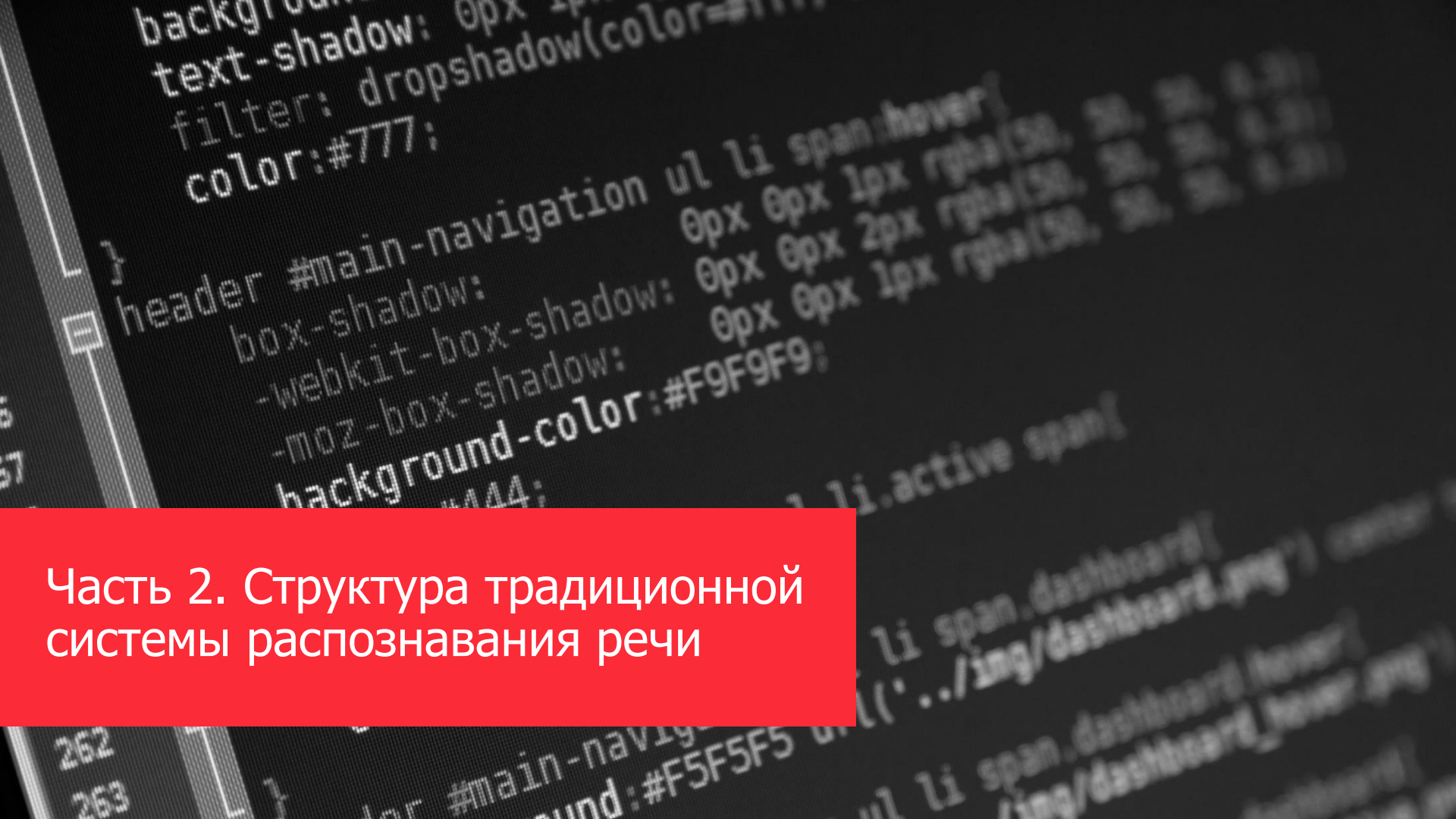
Блок 2. Автоматическое распознавание речи

Максим Кореневский

Старший научный сотрудник

ООО «ЦРТ-инновации», к.ф.-м.н.

Блок 2. Автоматическое распознавание речи (Automatic Speech Recognition, ASR)



Часть 2. Структура традиционной системы распознавания речи

Сравнение с эталоном (напоминание из лекции 1)

Достоинства и недостатки DTW-подхода в целом:

- \+ Интуитивность идеи
- \+ Простота реализации
- \+ Допустимо создавать эталоны не слов, а произвольных звуков
- – Необходимость хранить все эталоны
- – Ограниченность набора эталонов в смысле обобщающей способности
- – Невысокая точность
- – Маленький объем словаря

- Выход:
  - Создавать «модели» слов, описывающие все потенциальное множество их эталонов
  - Обучать модели по большим объемам данных
  - Для распознавания использовать только сами модели, без эталонов

План лекции

- Вероятностная постановка задачи распознавания речи
- Акустическая модель
- Языковая модель
- Лексикон
- Декодер
- Что требуется для создания системы распознавания речи?

План лекции

- Вероятностная постановка задачи распознавания речи
- Акустическая модель
- Языковая модель
- Лексикон
- Декодер
- Что требуется для создания системы распознавания речи?

Вероятностная постановка задачи распознавания

- Произнесена последовательность слов $W = (w_1, w_2, \dots, w_n)$
- По ней получена последовательность наблюдений $O = (o_1, o_2, \dots, o_T)$
- Как, зная O, найти W ?

$$W = \arg\max_W P(W|O) = \arg\max_W \frac{p(O|W)P(W)}{p(O)} = \arg\max_W p(O|W)P(W)$$

- За оценку правдоподобия $p(O|W)$ отвечает акустическая модель (классификатор)
- За оценку априорной вероятности $P(W)$ последовательности слов отвечает языковая модель
- За максимизацию всего произведения отвечает декодер

Акустическая модель

По последовательностям слов строить распределение $p(O|W)$ очень сложно

Акустическая модель

Разобьем фразу на отдельные слова – стало проще, т.к. слов конечное число

НО их по-прежнему слишком много (сотни тысяч)!

Акустическая модель

Каждое слово состоит из фонем. Фонем немного (~50) и они короткие
Давайте строить акустические модели на базе фонем!

Акустическая модель

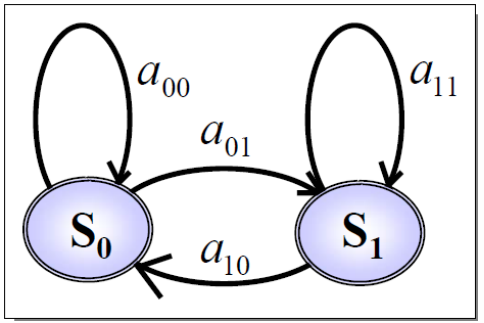
Марковский процесс с дискретным временем (цепь Маркова)

- Есть множество состояний $S = \{S_0, S_1, \dots, S_N\}$. В каждый момент процесс в одном из них.
- Переходы между состояниями недетерминированные.
- Марковское свойство (независимость от истории): $P(q_t = S_j | q_{t-1} = S_i) = f(S_i)$
- Начальные вероятности:

$$\pi_i = P(q_1 = S_i)$$

- Вероятности перехода:

$$a_{ij} = P(q_t = S_j | q_{t-1} = S_i), \qquad t = 2,3, \dots, T$$

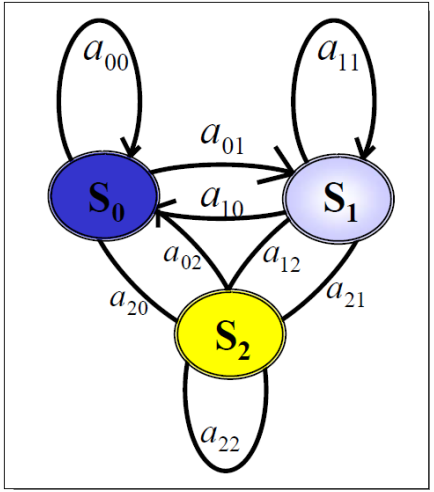


- Условия нормировки:

$$\sum_{i=0}^{N} \pi_i = 1, \qquad \sum_{j=0}^{N} a_{ij} = 1, \qquad i = 0,1, \dots, N$$

Акустическая модель

Пример: наблюдаемая Марковская цепь для погоды



- Три состояния: S_0 - дождь, S_1 - облака, S_2 - солнце
- Вероятности переходов заданы матрицей

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$

- Как рассчитать вероятность последовательности

  $O = \{\text{солнце}, \text{солнце}, \text{солнце}, \text{дождь}, \text{дождь}, \text{солнце}, \text{облака}, \text{солнце}\}$?
- Ей соответствует последовательность состояний $S = (2, 2, 2, 0, 0, 2, 1, 2)$.
- $P(O|\text{модель}) = P(S = (2, 2, 2, 0, 0, 2, 1, 2)|\text{модель}) =$

  $P(q_1 = 2)P(q_2 = 2|q_1 = 2) \cdots P(q_8 = 2|q_7 = 1) = \pi_2(0.8)^2 \cdot (0.1) \cdot (0.4) \cdot (0.3) \cdot (0.1) \cdot (0.2).$

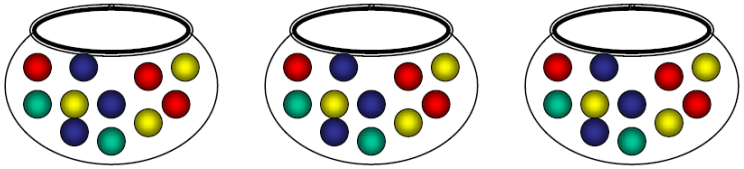
Акустическая модель

Скрытая Марковская модель (Hidden Markov Model, HMM)

- Наблюдаемая Марковская модель: $\lambda = (\pi, A)$, где $\pi_i = P(q_1 = i), \ \ i = 0,1, \dots, N$ - начальные вероятности, A – матрица переходов
- Скрытая (дискретная) Марковская модель: $\lambda = (\pi, A, B)$
  - Состояния процесса больше **не наблюдаются** непосредственно
  - Есть множество наблюдаемых значений V
  - В каждом состоянии задано вероятностное распределение наблюдаемых в нем значений:

    $V = \{v_1, v_2, \dots, v_M\}$ - множество значений, $b_{jk} = P(o_t = v_k | q_t = S_j)$ - наборы вероятностей,
  - B – набор этих вероятностных распределений: $B = \{b_0, b_1, \dots, b_N\}$
- Отдельные наблюдения при фиксированных состояниях НЕЗАВИСИМЫ (frame independence assumption)

Акустическая модель



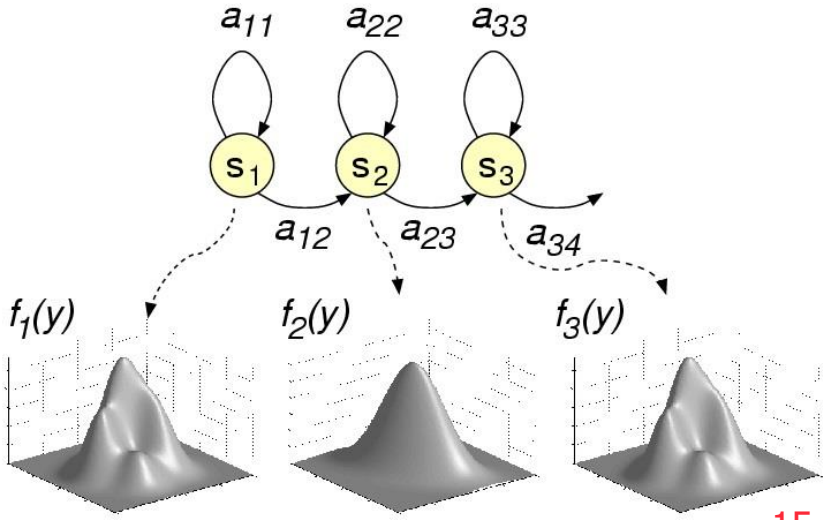
Пример с урнами и шарами

- Есть 3 урны, в каждой из которых определенное известное количество шаров красного, синего и зеленого цвета. И есть некий «аппарат»:
  1. Вначале аппарат выбирает урну наугад в соответствии с некими вероятностями π_i
  2. После этого аппарат достает из урны случайно выбранный шар и записывает его цвет
  3. Шар возвращается обратно в урну
  4. После этого аппарат выбирает, к какой урне переместиться согласно распределению a_{ij}
  5. Шаги 2-4 повторяются некоторое количество раз
- Наблюдатель видит только последовательность цветов, записанную аппаратом. Номера урн он не знает! Хочется уметь отвечать на вопросы:
  - Какова вероятность выбранной последовательности цветов?
  - Какой последовательности урн она наиболее вероятно соответствует?

Акустическая модель

Скрытая Марковская модель для речи

- Для речи $V = \mathbb{R}^d$, наблюдения $o_t \in V$ = векторы акустических признаков, состояния = части фонемы (начальная, средняя, финальная), но они от нас СКРЫТЫ!



Акустическая модель

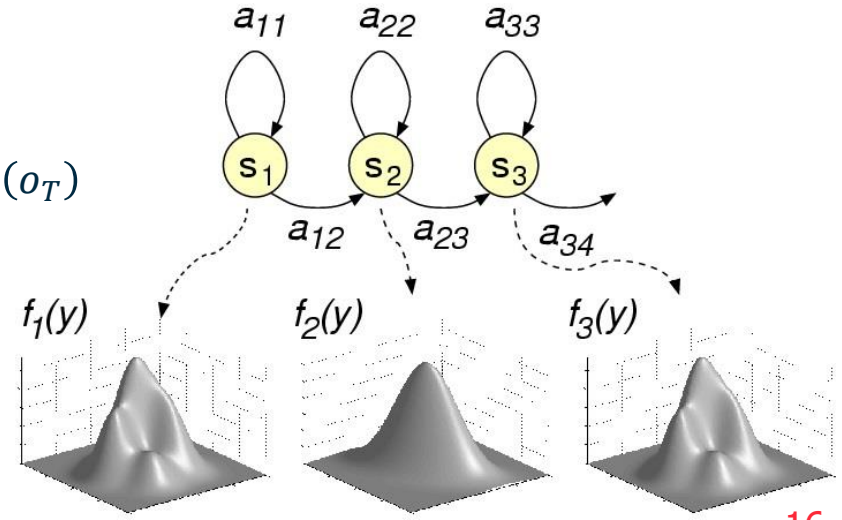
Скрытая Марковская модель для речи

- Для речи $V = \mathbb{R}^d$, наблюдения $o_t \in V$ = векторы акустических признаков, состояния = части фонемы (начальная, средняя, финальная), но они от нас СКРЫТЫ!
- Для данной последовательности состояний q:

$$P(q|\lambda) = \pi_{q_1} \cdot a_{q_1 q_2} \cdot a_{q_2 q_3} \cdots a_{q_{T-1} q_T}$$

$$P(O|q,\lambda) = \prod_{t=1}^{T} P(o_t|q_t,\lambda) = f_{q_1}(o_1) \cdot f_{q_2}(o_2) \cdots f_{q_T}(o_T)$$

$$P(O|\lambda) = \sum_q P(O,q|\lambda) = \sum_q P(O|q,\lambda) P(q|\lambda)$$



План лекции

- Вероятностная постановка задачи распознавания речи
- Акустическая модель
- Языковая модель
- Лексикон
- Декодер
- Что требуется для создания системы распознавания речи?

Языковая модель (Language model, LM)

- Языковая модель оценивает вероятность последовательности слов $P(W)$
- По теореме умножения вероятностей:

$$P(W) = P(w_1 w_2 \dots w_L) = P(w_1)P(w_2|w_1)P(w_3|w_1 w_2) \cdots P(w_L|w_1 w_2 \dots w_{L-1})$$

$$= P(w_1) \prod_{i=2}^{L} P(w_i|w_1 w_2 \dots w_{i-1})$$

- Итак, основная задача – научиться вычислять $P(w_i|w_1 w_2 \dots w_{i-1})$.
- Проблемы:
  - Последовательности могут быть произвольной длины
  - Чем длиннее «история» тем меньше шансов, что такое есть в обучающих данных
  - Некоторые, даже короткие, последовательности вообще никогда не встречаются в языке

Языковая модель

Статистическая n-граммная (n-gram) языковая модель:

- Ограничим максимально возможную длину истории $(n-1)$ словом:

$$P(w_i|w_1 w_2 \dots w_{i-1}) \approx P(w_i|w_{i-n+1} w_{i-n+2} \dots w_{i-1})$$

- Последовательности из n слов $w_{i-n+1} w_{i-n+2} \dots w_{i-1} w_i$ называются **n-граммы** (n-gram):
  - $n = 1$ – униграммы, $n = 2$ – биграммы, $n = 3$ – триграммы и т.д.
- В результате, например, для триграммной модели:

$$P(\text{"мой дядя самых честных правил"}) =$$
$$P(\text{"мой"}) \cdot P(\text{"дядя"}|\text{"мой"}) \cdot P(\text{"самых"}|\text{"мой дядя"}) \cdot$$
$$\cdot P(\text{"честных"}|\text{"дядя самых"}) \cdot P(\text{"правил"}|\text{"самых честных"})$$

- Поэтому в триграммной модели должны присутствовать не только вероятности триграмм, но также вероятности биграмм и униграмм

Языковая модель

Оценка вероятностей n-грамм:

- Пусть дан большой обучающий корпус.
- Подсчитаем статистику количества появлений всевозможных n-грамм в нем.
- Тогда оценка максимального правдоподобия для вероятности имеет вид

$$P(w_i|w_{i-n+1}w_{i-n+2}\dots w_{i-1}) \approx \frac{\#\text{ появлений } w_{i-n+1}w_{i-n+2}\dots w_{i-1}w_i}{\#\text{ появлений } w_{i-n+1}w_{i-n+2}\dots w_{i-1}}$$

- Т.е. мы вычисляем ЧАСТОТУ появления обучающем корпусе данной последовательности слов среди всех последовательностей С ТОЙ ЖЕ ИСТОРИЕЙ и разными последними словами!
- Обычно в модели сохраняются логарифмы, это позволяет перейти от произведений к суммам и избежать потери точности.

Языковая модель

Оценка качества языковой модели:

- Качество ЯМ определяется тем, насколько хорошо она способна предсказывать очередное слово по его истории. Т.е. насколько высокие вероятности она дает РЕАЛЬНЫМ предложениям.
- Перплексия (perplexity) – мера точности ЯМ:

$$PPL(w_1 w_2 \dots w_L) = P(w_1 w_2 \dots w_L)^{-1/L} = e^{-\frac{\log P(w_1 w_2 \dots w_L)}{L}}$$

- Перплексия всегда больше единицы, чем меньше перплексия, тем лучше ЯМ
- Важно:
  - Вычислять перплексию на отдельном тексте, не входящем в обучающую выборку
  - Сравнивать разные модели по перплексии на одном и том же тексте (а не на разных)

Языковая модель

Discounting

- Для размера словаря в 1000 слов число различных триграмм – миллиард
- НО: большинство триграмм вообще никогда не встречаются в речи
- А часть триграмм в речи есть, но их может не быть в обучающем корпусе (unseen), на них надо бы выделить некоторую долю вероятности
- Для этого используется discounting (уменьшение вероятности встреченных n-грамм):

$$P(w_i|w_{i-n+1}w_{i-n+2}\dots w_{i-1}) \approx \frac{D(\#\text{ появлений } w_{i-n+1}w_{i-n+2}\dots w_{i-1}w_i)}{\#\text{ появлений } w_{i-n+1}w_{i-n+2}\dots w_{i-1}}$$

- Good-Turing discounting: $D_{GT}(C) = (C+1)\frac{N_{C+1}}{N_c}$,
- Kneser-Ney discounting: $D_{KN} = \frac{C_{KN}(w_{i-n+1}w_{i-n+2}\dots w_{i-1}w_i)}{\sum_W C_{KN}(w_{i-n+1}w_{i-n+2}\dots w_{i-1}w)}$

Языковая модель

Сглаживание, откаты (back-off):

- Как распределить discounted вероятность между unseen n-граммами?
- Один из наиболее распространенных подходов: использовать так называемые веса «отката» (back-off weights), Katz, 1987:

$$P_{smooth}(w_3|w_1w_2) = \begin{cases} P(w_3|w_1w_2), & \text{если } \#(w_1w_2w_3) > 0 \\ \alpha(w_1w_2)P_{smooth}(w_3|w_2), & \text{если } \#(w_1w_2) > 0 \\ P_{smooth}(w_3|w_2), & \text{в противном случае} \end{cases}$$

- Для вероятностей, входящих в правую часть используется то же правило:

$$P_{smooth}(w_3|w_2) = \begin{cases} P(w_3|w_2), & \text{если } \#(w_2w_3) > 0 \\ \alpha(w_2)P(w_3), & \text{в противном случае} \end{cases}$$

- Формат ARPA LM: для всех n-грамм не высшего порядка хранятся $\log P_{smooth}$ и $\log \alpha$

Языковая модель

Пример языковой модели в формате ARPA:

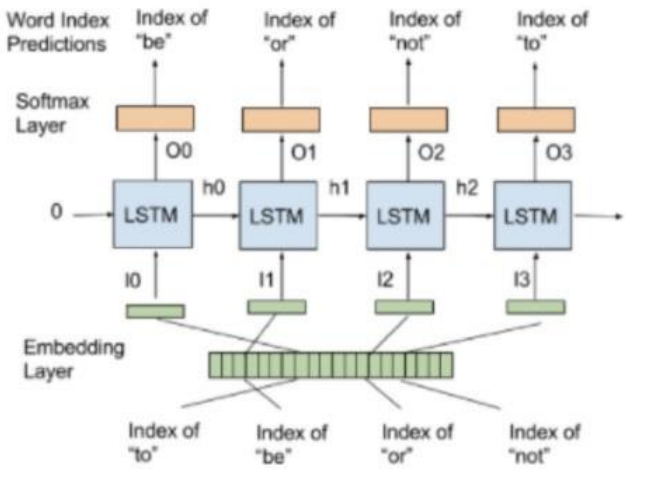
```
\data\
ngram 1=37445
ngram 2=138797
ngram 3=51837
ngram 4=53201

\1-grams:
-1.727079 A    -1.184703
-5.57354  AACHEN    -0.30103
-5.57354  AAMI -0.30103
-4.833177 AARON     -0.3245111
-5.27251  AARONS    -0.39794
-5.57354  AARRON    -0.30103
…
-0.4887864      INCLUDE A -0.8016323
-1.698039       INCLUDE ART
-1.698428       INCLUDE BODY
…
-0.1107919      ALSO INCLUDE A -0.4771213
-0.1641603      TO INCLUDE A   -1.079181
-0.611348       TO INCLUDE EVERY    -0.69897
…
-0.153755       ALSO INCLUDE A TEN
-0.01548983     TO INCLUDE A NEW
-0.102658       WILL INCLUDE A SHIFT
-0.01771067     IT INCLUDES A STOPOVER
…
\end\
```

Языковая модель

Языковые модели на основе нейронных сетей:

- На вход принимают последовательность индексов слов в словаре (или 1-hot векторы)
- Предсказывают вероятности ВСЕХ слов словаря при заданной истории ОДНОВРЕМЕННО
- Обучение: непосредственная минимизация перплексии на обучающем корпусе текстов



Языковая модель

Языковые модели на основе нейронных сетей, за и против:

- Плюсы:
  - Способны учитывать значительно более длинный контекст
  - Как правило, обеспечивают значительно меньшую перплексию, чем n-gram LM
  - Способны генерировать более естественный текст
  - В результате обучения получаются полезные представления слов (word2vec)
  - Могут обучаться на очень больших корпусах данных
  - Могут дообучаться для решения различных задач NLP
- Минусы
  - Размер выходного слоя равен размеру словаря!
  - Языковые модели на основе рекуррентных нейронных сетей хранят «историю» во внутреннем состоянии модели.

План лекции

- Вероятностная постановка задачи распознавания речи
- Акустическая модель
- Языковая модель
- Лексикон
- Декодер
- Что требуется для создания системы распознавания речи?

Лексикон

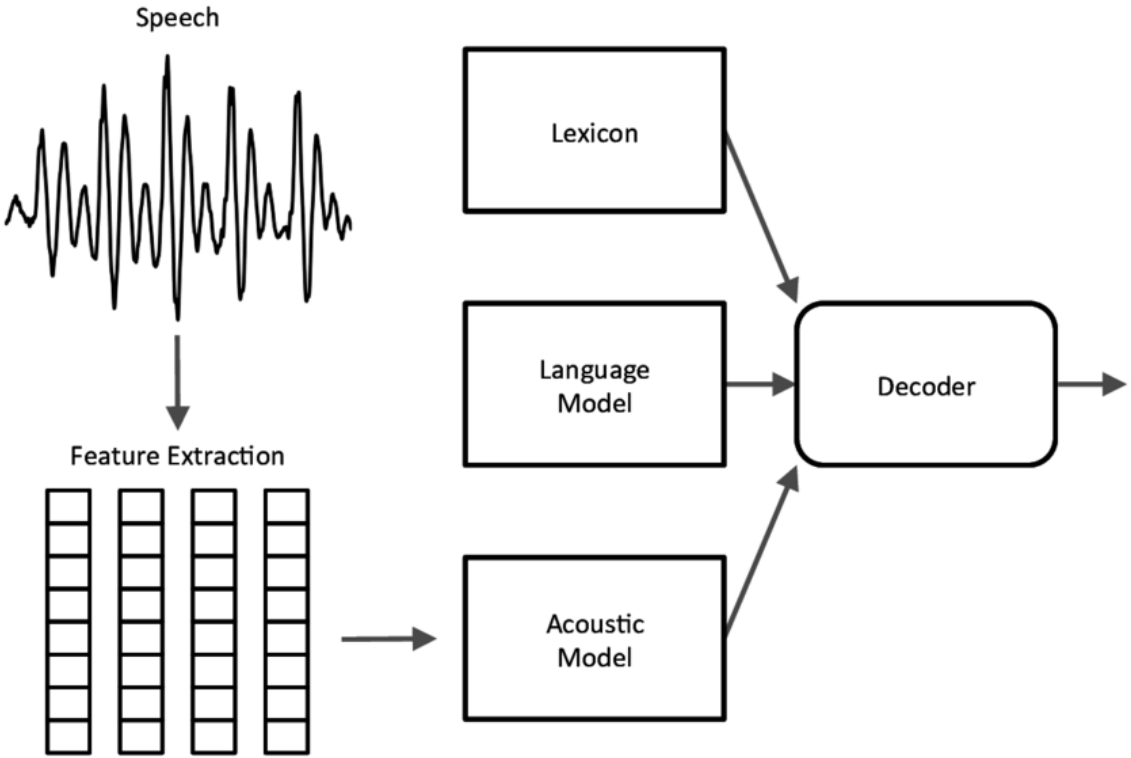
Лексикон = словарь транскрипций (**орфоэпический** словарь)

- Запись слов последовательностью фонем:
  - мама m a0 m a4
  - мыла m y0 l a4
  - морковь m a1 r k o0 f’
- Есть слова с множественными транскрипциями (омографы): зАмок-замОк
- Составление словаря: работа лингвистов, дорого, долго
- Альтернатива: G2P (grapheme-to-phoneme) конвертер (Sequitur, Phonetisaurus)
  - Обучается на небольшом словаре, составленном вручную
  - Способен генерировать варианты транскрипции для незнакомых слов и приписывать им вероятности
- Есть системы, работающие непосредственно с графемами, им лексикон не нужен

План лекции

- Вероятностная постановка задачи распознавания речи
- Акустическая модель
- Языковая модель
- Лексикон
- Декодер
- Что требуется для создания системы распознавания речи?

Напоминание: архитектура (традиционной) ASR-системы



Декодер

- Задача декодера – имея последовательность наблюдений O, найти последовательность слов W:

$$W = \arg\max_W p(O|W)P(W) = \arg\max_W [\ln p(O|W) + \ln P(W)]$$

- Иногда рассматривают более чуть общую задачу:

$$W = \arg\max_W [\ln p(O|W) + \beta \ln P(W)]$$

- У нас нет акустической модели для оценки $p(O|W)$ на всей последовательности слов, но, допустим, есть HMM для отдельных фонем.
- Используя лексикон, можно построить HMM для любого слова, а значит и для любой последовательности слов!
- Т.е. (теоретически) можно построить огромную HMM и искать наилучший путь по ней!
- Эту идею мы обсудим подробнее в следующей лекции.

План лекции

- Вероятностная постановка задачи распознавания речи
- Акустическая модель
- Языковая модель
- Лексикон
- Декодер
- Что требуется для создания системы распознавания речи?

Сбор и подготовка данных для обучения

- Подготовка акустической базы (для обучения акустической модели):
  - Запись фонограмм / поиск и скачивание аудиоданных в свободном доступе / покупка речевой базы
  - Предобработка фонограмм (разделение каналов стерео, шумоочистка, нарезка)
  - Подготовка эталонных текстовок (если они отсутствуют)
  - Аннотирование речевых данных (диктор, канал записи, особенности записи и т.д.)
- Подготовка текстовой базы (для обучения языковой модели):
  - Набор текстовых данных из различных источников (книги, фильмы, телефонные разговоры, социальные сети, Википедия, TV-программы, выпуски новостей и т.д.)
  - Парсинг и фильтрация данных (удаление html-тегов, повторов, рекламы и т.д.)
  - Нормализация текстов (регистр, кодировка, раскрытие числительных, аббревиатур и т.д.)

Спасибо за внимание!

Вопросы?