



Natural Language Processing

Language models



Проверить, идет ли запись

Меня хорошо видно && слышно?



Ставим "+", если все хорошо
"-", если есть проблемы



Тема вебинара

Language models



Никита Мартынов

Head NLP Engineer @ SberDevices, Team Lead @ MIL Team

(Ex) Data Scientist / NLP Engineer @ Tinkoff, MTS AI, Financial Technologies Lab @ MIPT

tg: @go_bobert



Правила вебинара



Активно
участвуем



Off-topic обсуждаем
в Slack #канал группы
или #general

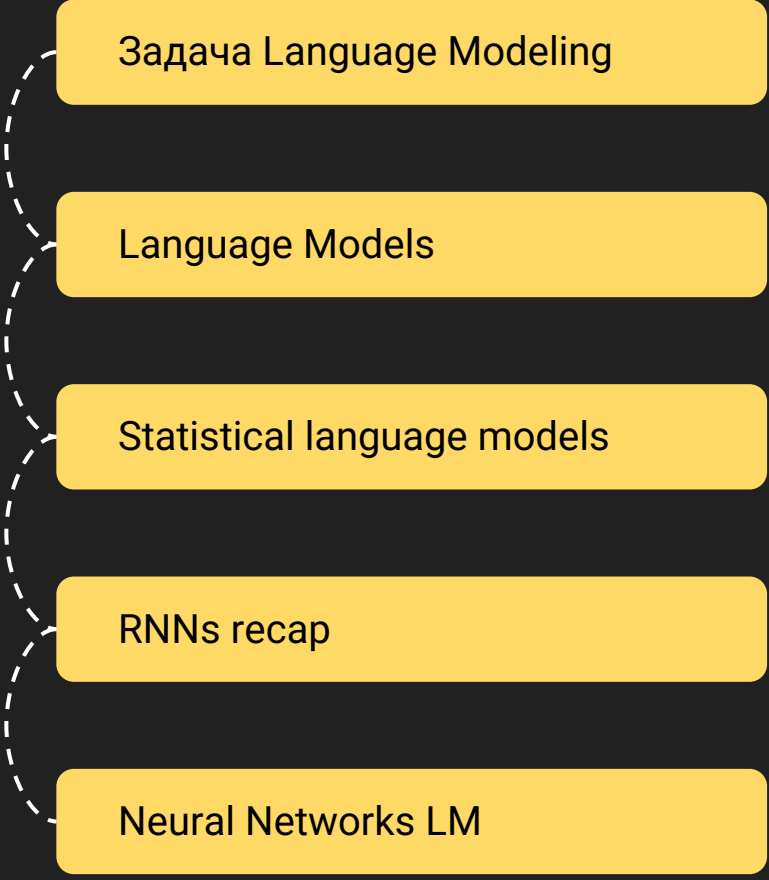


Задаем вопрос
в чат или ГОЛОСОМ



Вопросы вижу в чате,
могу ответить не сразу

Маршрут вебинара



Задача Language Modeling

Language Models

Statistical language models

RNNs recap

Neural Networks LM



Задача Language Modeling

Задача Language Modeling

У нас могут быть различные языковые сущности:

1. Символы / буквы;
2. Слова;
3. Части слов;
4. Предложения;
5. Части предложения;
6. Абзацы;
7. Параграфы;
8. Любой текст;

Задача: оценить вероятность этих языковых сущностей!

Задача Language Modeling

Как оценивать вероятность текста?

1. Рассматриваем предложение / слово / любую языковую структуру как последовательность;
2. Разбиваем последовательность на элементы;
3. У нас получается конечное множество (мы всегда можем ограничить словарь);
4. Давайте определим на подмножествах этого множества вероятностную функцию...

Математическая пауза

Как правильно определять вероятности:

1. Пусть есть множество элементов \mathcal{U} , оно может быть конечным, может быть счётным, может быть несчётным;
2. Определим на этом множестве систему его подмножеств \mathcal{A} , такую что:
 - a. $\mathcal{U} \in \mathcal{A}$;
 - b. $A \in \mathcal{A}, B \in \mathcal{A} \rightarrow A \cup B \in \mathcal{A}, A \cap B \in \mathcal{A}$;
 - c. $A \in \mathcal{A} \rightarrow \bar{A} \in \mathcal{A}$;
3. Это называется алгеброй – система подмножеств, замкнутой относительно конечного числа операций объединения, пересечения и взятия дополнения;
4. **БОНУС:** алгебра, замкнутая относительно счётного числа пересечений и объединений, называется σ -алгеброй;
5. Пару $\langle \mathcal{U}, \mathcal{A} \rangle$ называют измеримым пространством;
6. Теперь заведём вероятность на множествах из \mathcal{A} :
 - a. $P(A) \geq 0 \forall A \in \mathcal{A}$;
 - b. $P(\mathcal{U}) = 1$;
 - c. $P(\sum_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} P(A_n), A_i A_j = \emptyset \forall i, j: i \neq j$;
7. Тройка $\langle \mathcal{U}, \mathcal{A}, P \rangle$ называется вероятностным пространством;

Задача Language Modeling

Что там насчёт вероятностных функций на пространстве языковых конструкций?

1. \mathcal{U} - это будет как раз множество всех элементов в словаре (слова, символы, и тд);
2. Совокупности «слов» будут как раз алгеброй \mathcal{A} :

- a. $\mathcal{U} \in \mathcal{A}$;
- b. $A \in \mathcal{A}, B \in \mathcal{A} \rightarrow A \cup B \in \mathcal{A}, A \cap B \in \mathcal{A}$;
- c. $A \in \mathcal{A} \rightarrow \bar{A} \in \mathcal{A}$;

3. А как тогда считать P ?

Задача Language Modeling

Как посчитать P?

1. Пусть предложение будет $S = [s_1, \dots, s_n]$;

2. $P(S) = P(s_1, \dots, s_n)$;

3. Chain rule:

$$\begin{aligned} P(S) &= P(s_1, \dots, s_n) = P(s_1, \dots, s_{n-1})P(s_n | s_1, \dots, s_{n-1}) = \\ &P(s_1, \dots, s_{n-2})P(s_{n-1} | s_1, \dots, s_{n-2})P(s_n | s_1, \dots, s_{n-1}) = P(s_1)P(s_1 | s_2) \dots P(s_n | s_1, \dots, s_{n-1}) \\ &= \prod_{i=1}^n P(s_i | s_j, j < i); \end{aligned}$$

4. Как посчитать тогда $P(s_i | s_j, j < i)$?

$$P(s_i | s_j, j < i) = \frac{\text{counter}(s_1, \dots, s_i)}{\text{counter}(s_1, \dots, s_{i-1})}$$

Слишком разряженная статистика получится, что тогда делать?

Задача Language Modeling

Марковское свойство!

$$P(X_n = x_n | X_n = x_n, \dots, X_1 = x_1) = P(X_n = x_n | X_{n-1} = x_{n-1})$$

Слишком строго, давайте расслабим:

$$P(X_n = x_n | X_n = x_n, \dots, X_1 = x_1) = P(X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_{n-k} = x_{n-k})$$

Тогда получаем:

$$P(s_i | s_j, j < i) = P(s_i | s_j, j < i, j \geq i - k)$$

$$P(S) = \prod_{i=1}^n P(s_i | s_j, j < i, j \geq i - k);$$

Задача Language Modeling

Где это может применяться:

1. Сервисы спелл-чека;
2. Поисковые движки;
3. Keyboard Autocomplete;
4. Чат-боты, голосовые / текстовые ассистенты;
5. Машинный перевод;

В целом любая задача текстовой генерации;

Задача Language Modeling

Почему это может плохо работать:

1. Нули: слишком разреженная статистика;
2. Чтобы нормально что-то оценить, нужен огромный корпус;
3. Вычисления пропорциональны длине контекста и размеру корпуса;
4. По факту нельзя использовать длинный контекст;

Neural LMs

Neural LMs

Как оценивать вероятность предложения:

1. В целом концепция, которую мы предложили, никуда не уходит:

$$\begin{aligned} P(S) &= P(s_1, \dots, s_n) = P(s_1, \dots, s_{n-1})P(s_n | s_1, \dots, s_{n-1}) = \\ &P(s_1, \dots, s_{n-2})P(s_{n-1} | s_1, \dots, s_{n-2})P(s_n | s_1, \dots, s_{n-1}) = P(s_1)P(s_1 | s_2) \dots P(s_n | s_1, \dots, s_{n-1}) \\ &= \prod_{i=1}^n P(s_i | s_j, j < i); \end{aligned}$$

2. Сделать репрезентацию контекста $s_j, j < i$;
3. По ней предиктить следующее слово;

Neural LMs

Получается пайплайн:

1. Разбить текст на токены (слова, символы, n-grams);
2. Векторная репрезентация токенов;
3. Получить репрезентацию контекста;
4. По репрезентации контекста предсказать следующее слово – вероятностное распределение на словаре;
5. Используем лосс:

$$L = - \sum_{t=1}^n \log p(y_t | x_1, \dots, x_t)$$

RNNs recap

RNNs recap

Информацию будем хранить в стейтах – векторах;

Введём обозначения:

1. Вектор, который хранит информацию о предыдущих шагах, - h_t ;
2. Вектор с новой информацией - x_t ;

Обновление вектора состояния:

$$h_t = f(h_t, x_t)$$

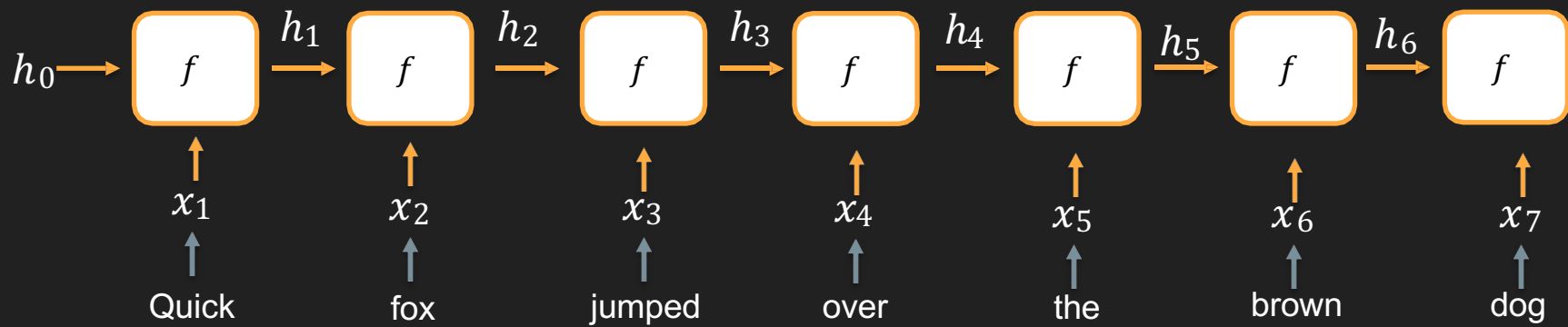
Параметризуем модель:

$$h_t = f(h_t, x_t; \theta)$$

Назовем это Recurrent Neural Network

Устройство RNN Cell

Как это выглядит:



Как устроена f :

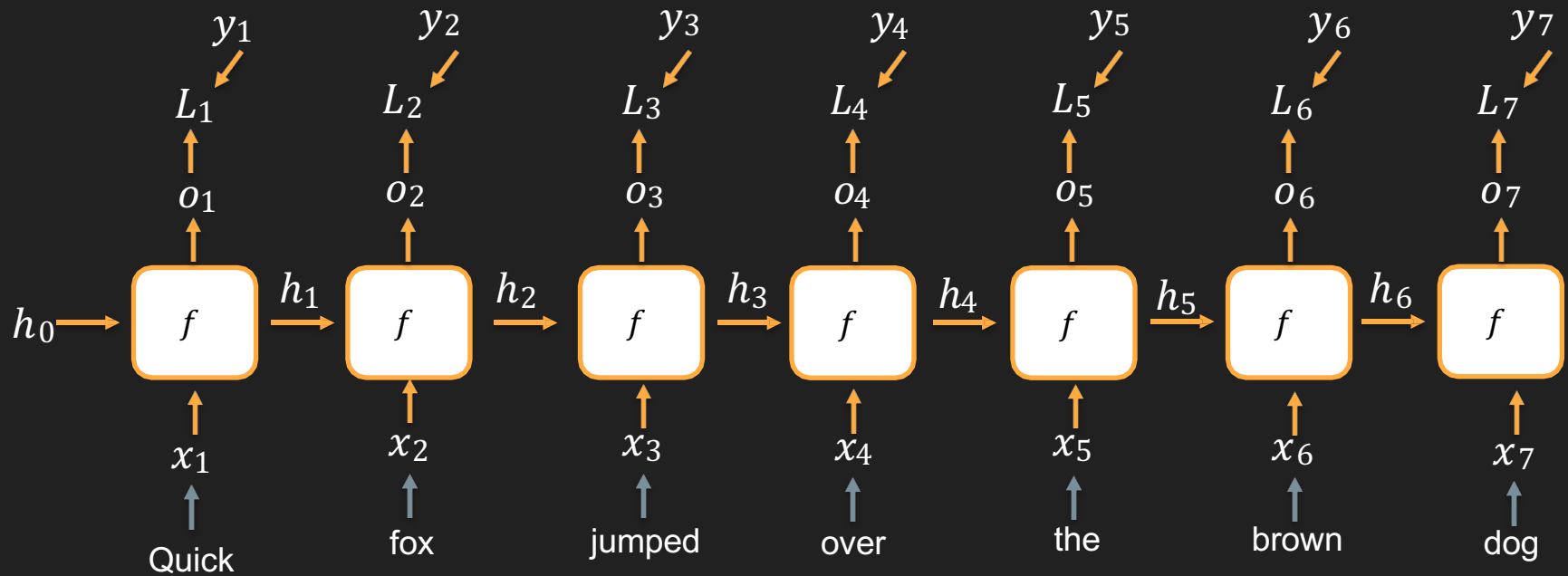
1. θ — это множество весов, пусть это будут W, U, V ;
2. $a_t = Wh_t + Ux_t + b$;
3. $h_t = \tanh(a_t)$;

Если хотим что-то предсказывать на каждом шаге:

1. $o_t = Vh_t + c$;
2. $y_t = \text{SM}(o_t)$

BBTT

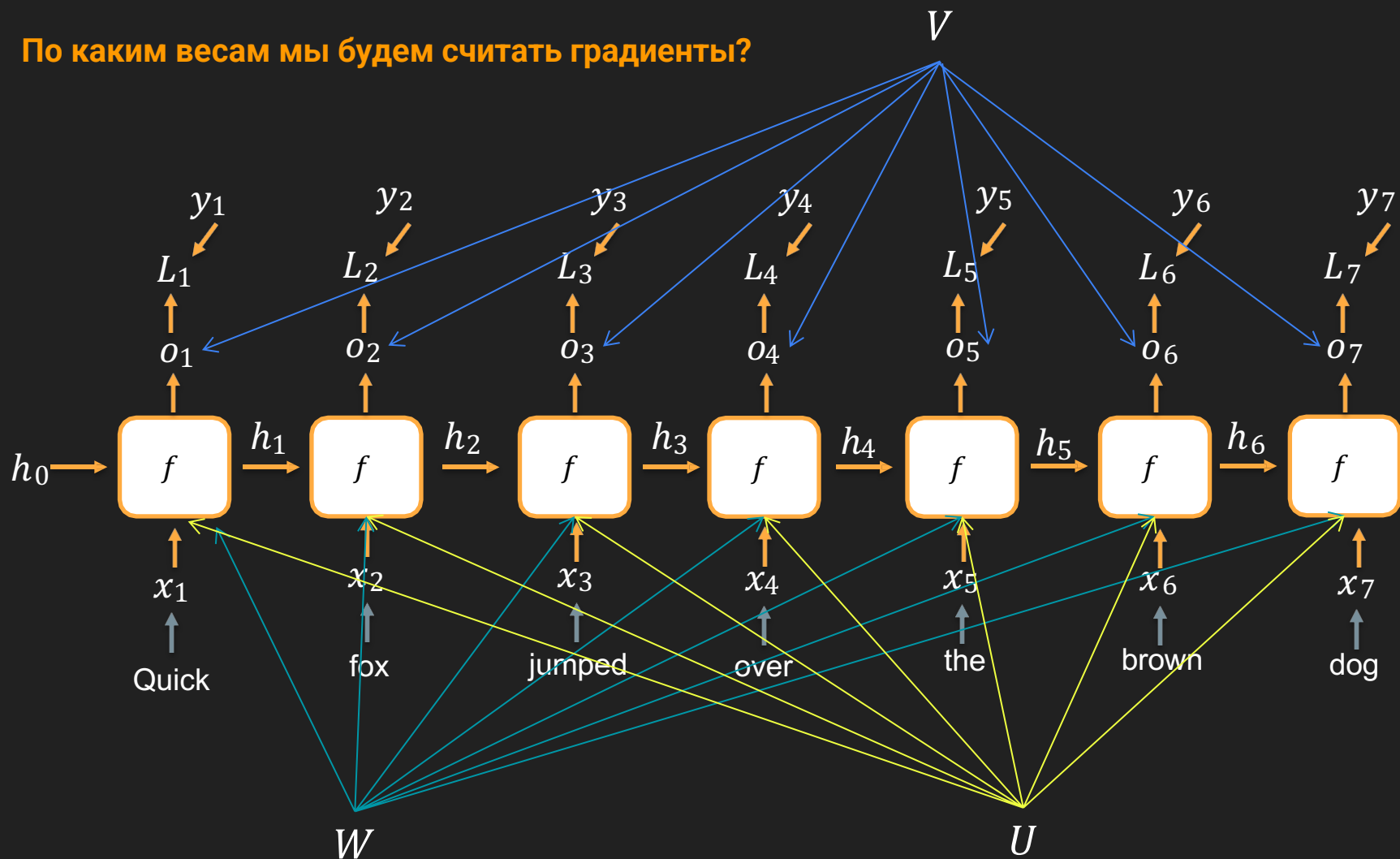
Back-Propagation Through Time



$$L = - \sum_{t=1}^n \log p(y_t | x_1, \dots, x_t)$$

ВРТТ

По каким весам мы будем считать градиенты?



Везде информацию переносит h_t , поэтому ключевой вопрос, как брать градиенты по нему

LSTM, GRU

$$z_t = [h_{t-1}, x_t]$$

$$f_t = \sigma(W_f \cdot z_t + b_f)$$

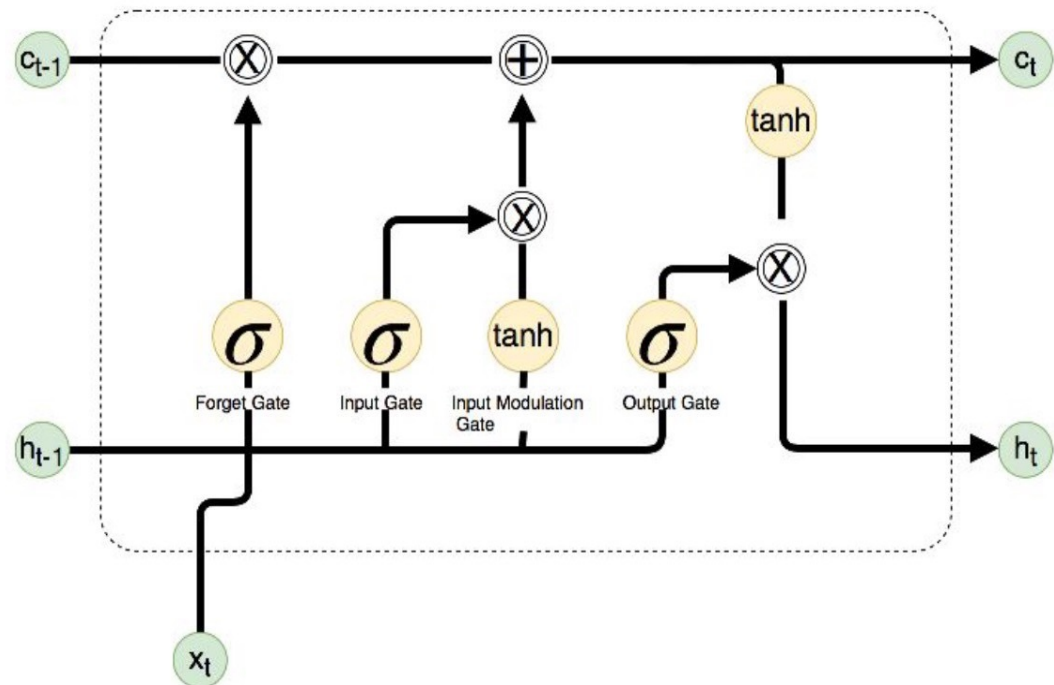
$$i_t = \sigma(W_i \cdot z_t + b_i)$$

$$\hat{C}_t = \tanh(W_c \cdot z_t + b_c)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \hat{C}_t$$

$$o_t = \sigma(W_o \cdot z_t + b_o)$$

$$h_t = o_t \cdot \tanh(C_t)$$



LSTM, GRU

Update gate: controls what parts of hidden state are updated vs preserved

Reset gate: controls what parts of previous hidden state are used to compute new content

New hidden state content: reset gate selects useful parts of prev hidden state. Use this and current input to compute new hidden content.

Hidden state: update gate simultaneously controls what is kept from previous hidden state, and what is updated to new hidden state content

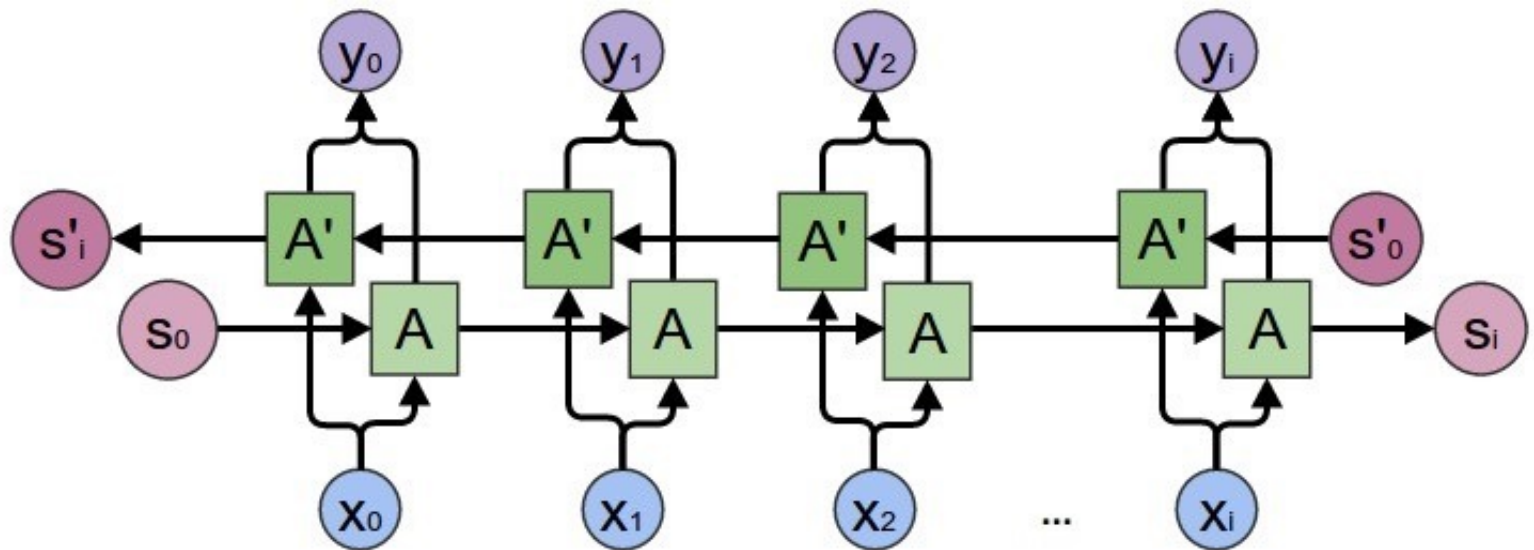
$$\mathbf{u}^{(t)} = \sigma \left(\mathbf{W}_u \mathbf{h}^{(t-1)} + \mathbf{U}_u \mathbf{x}^{(t)} + \mathbf{b}_u \right)$$

$$\mathbf{r}^{(t)} = \sigma \left(\mathbf{W}_r \mathbf{h}^{(t-1)} + \mathbf{U}_r \mathbf{x}^{(t)} + \mathbf{b}_r \right)$$

$$\tilde{\mathbf{h}}^{(t)} = \tanh \left(\mathbf{W}_h (\mathbf{r}^{(t)} \circ \mathbf{h}^{(t-1)}) + \mathbf{U}_h \mathbf{x}^{(t)} + \mathbf{b}_h \right)$$

$$\mathbf{h}^{(t)} = (1 - \mathbf{u}^{(t)}) \circ \mathbf{h}^{(t-1)} + \mathbf{u}^{(t)} \circ \tilde{\mathbf{h}}^{(t)}$$

LSTM, GRU

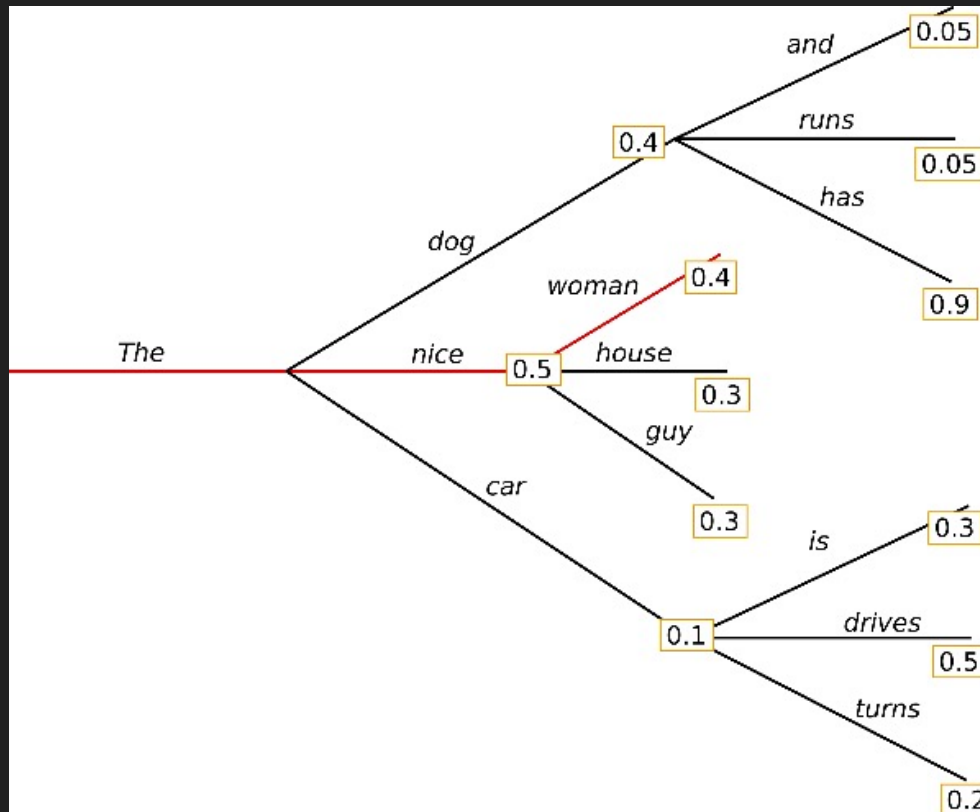


Text Generation

Text Generation

Есть три основные стратегии:

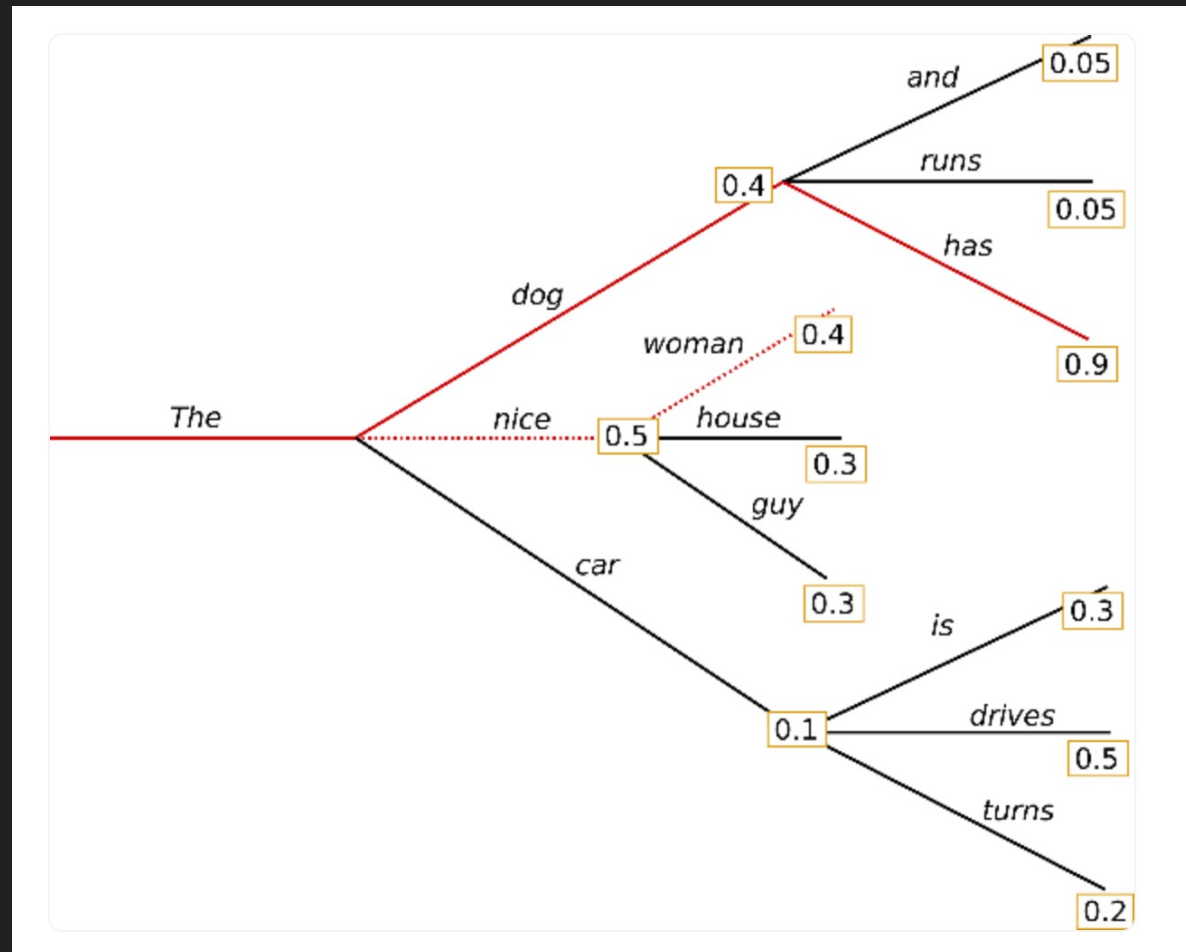
1. Grid search;



Text Generation

Есть три основные стратегии:

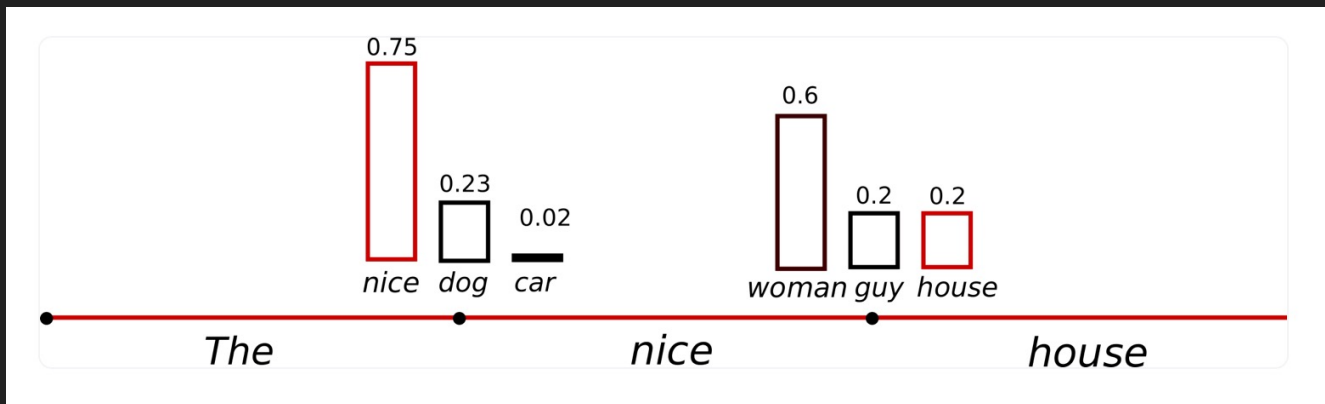
1. Beam search;



Text Generation

Есть три основные стратегии:

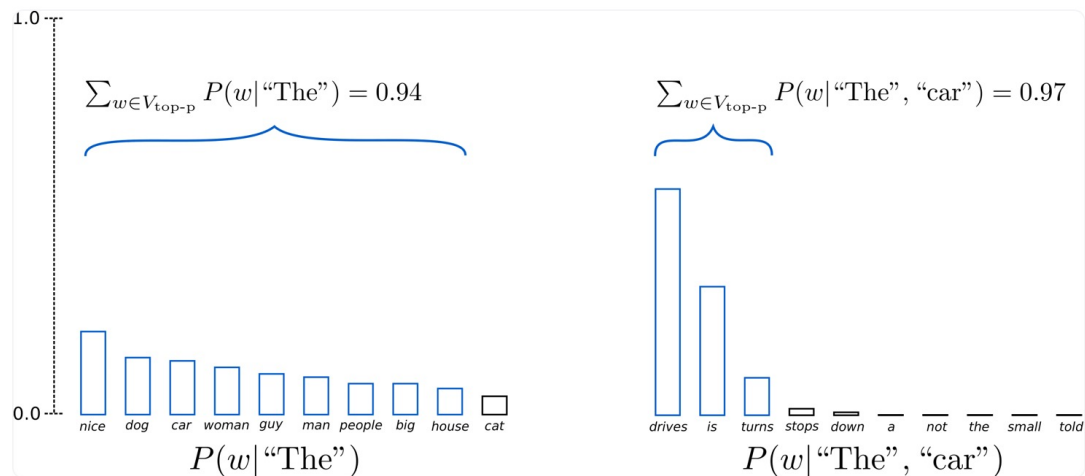
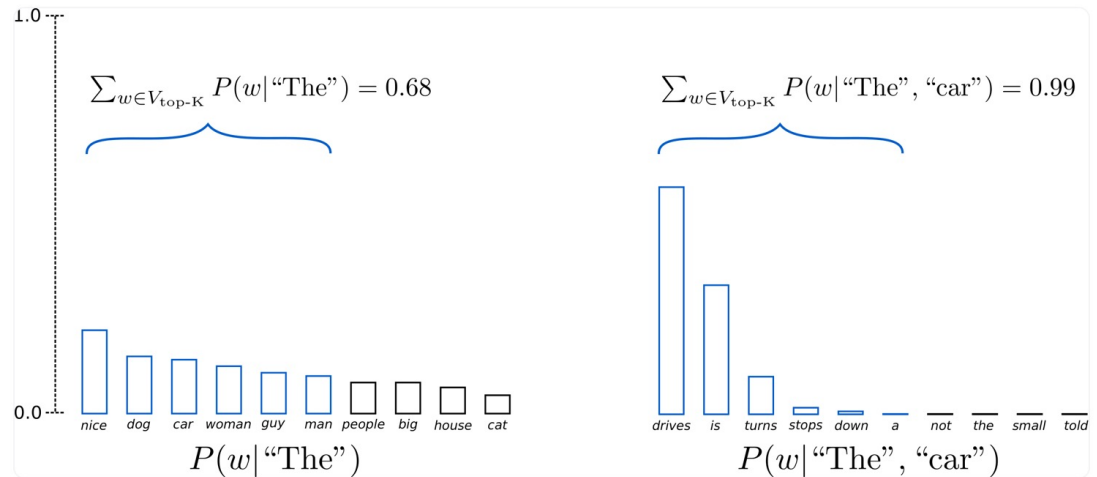
1. Sampling;



Text Generation

Есть три основные стратегии:

1. Sampling;



Evaluation

Evaluation

Как валидировать модели;

1. Перплексия:

$$PPL(W) = P(w_1, w_2, \dots, w_N)^{-\frac{1}{N}} = \sqrt[N]{1/P(w_1, w_2, \dots, w_N)}$$

2. Чем ниже, тем лучше;

3. Downstream задачи:

- a) Spellchecking;
- b) NMT;
- c) Text generation;
- d) etc.;

**Заполните, пожалуйста,
опрос о занятии
по ссылке в чате**