ТЕХНОСФЕРА

# Лекция 10
# Нейронные сети для обработки естественного языка

Полыковский Даниил

10 апреля 2016 г.

# Задачи NLP

- Машинный перевод
- Анализ тональности
- Чат-боты
- Понимание естественного языка
- Понимание изображений

# Entity tracking

| |
|---|
| mary got the milk there<br>john moved to the bedroom<br>sandra went back to the kitchen<br>mary travelled to the hallway<br>john got the football there<br>john went to the hallway<br>john put down the football<br>mary went to the garden<br>john went to the kitchen<br>sandra travelled to the hallway<br>daniel went to the hallway<br>mary discarded the milk<br>where is the milk ? |
| answer: garden |

# Visual QA[1]

Who is wearing glasses?

man

woman

Where is the child sitting?

fridge

arms

Is the umbrella upside down?

yes

no

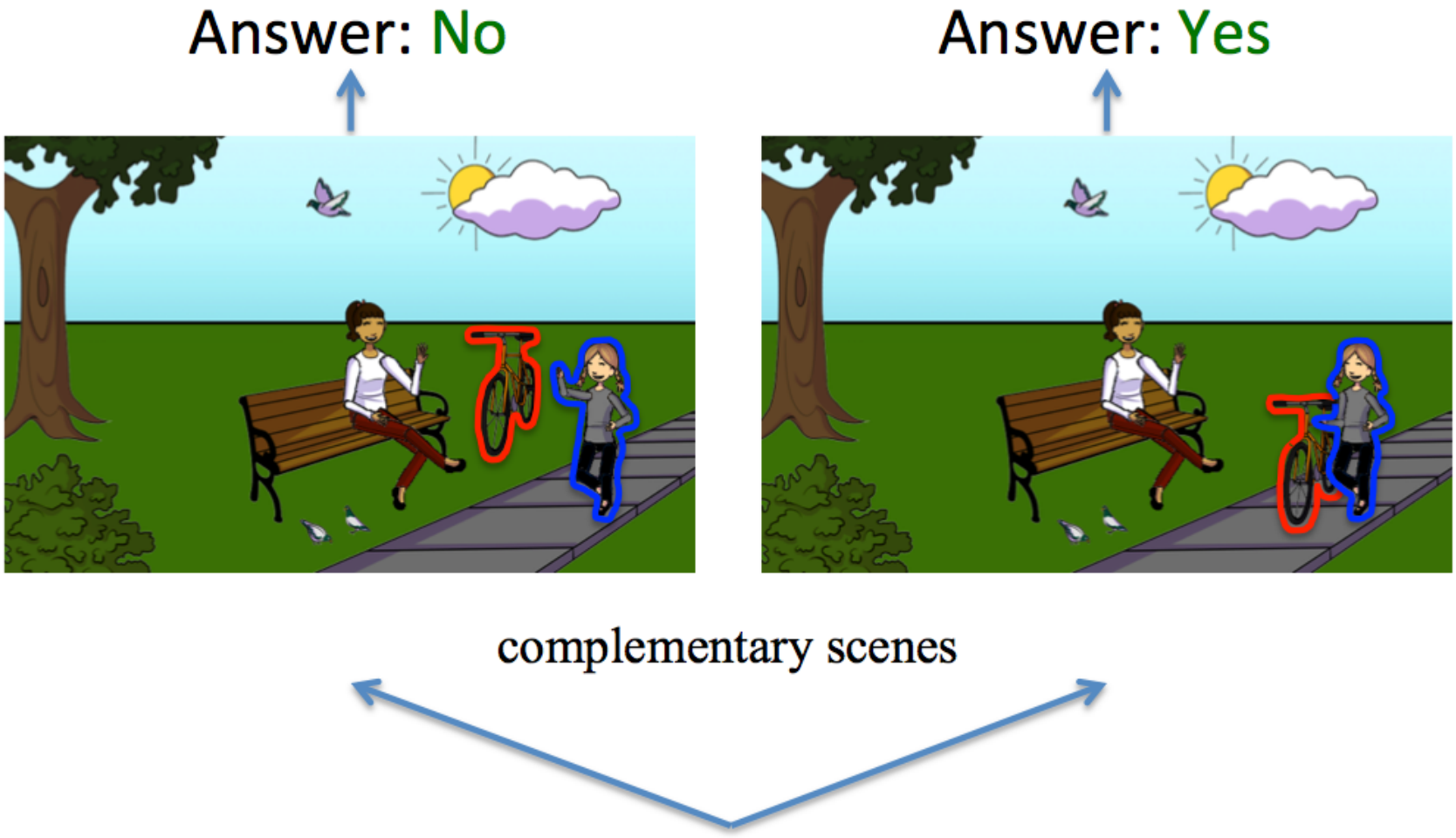
How many children are in the bed?

2

1

[1]Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering (CVPR 2017)

# Visual QA[2]




[2]Yin and Yang: Balancing and Answering Binary Visual Questions (CVPR 2016)

# Анализ тональности

## Отзыв положительный или отрицательный?

y=1 Мне очень понравился этот фильм. Никогда раньше ничего подобного не видел!!!)))

y=0 Ужасно! Ушла с середины фильма, тк больше невозможно было смотреть.

y=0 Ну да, конечно. Просто отличный фильм.

# Анализ тональности

## Отзыв положительный или отрицательный?

y=1 Мне очень понравился этот фильм. Никогда раньше ничего подобного не видел!!!)))

y=0 Ужасно! Ушла с середины фильма, тк больше невозможно было смотреть.

y=0 Ну да, конечно. Просто отличный фильм.

Простой подход: Bag-of-words + Logistic regression Какие есть проблемы у такого подхода?

- Не учитывает сарказм
- Не учитывает схожесть слов (например, кот $\leftrightarrow$ котенок)
- Не учитывает порядок слов

# Представление слов

## Задача

Сопоставить каждому слову $w$ из словаря $V$ вектор $e(w)$.

Подходы:

- One-hot encoding
- CW
- CBOW
- Skip-grams

# One-hot encoding

Кодируем слово $w_i$ вектором $[0, 0, \ldots, 0, \underbrace{1}_{i}, 0, \ldots, 0]^T$ Плюсы:

- Просто реализовать
- Можно использовать разреженное представление

Минусы:

- Не учитывает близость слов
- Огромная размерность

# Counts

... and the cute kitten purred and then ...
... the cute furry cat purred and miaowed ...
... that small kitten miaowed and she ...
... the loud furry dog ran and bit ...

Словарь: bit, cute, furry, loud, miaowed, purred, ran, small

kitten: cute, purred, small, miaowed $\Rightarrow [0, 1, 0, 0, 1, 1, 0, 1]^T$
cat: cute, furry, miaowed $\Rightarrow [0, 1, 1, 0, 1, 0, 0, 0]^T$
dog: loud, furry, ran, bit $\Rightarrow [1, 0, 1, 1, 0, 0, 1, 0]^T$

$$sim(w_1, w_2) = \frac{<w_1, w_2>}{||w_1|| \cdot ||w_2||}$$

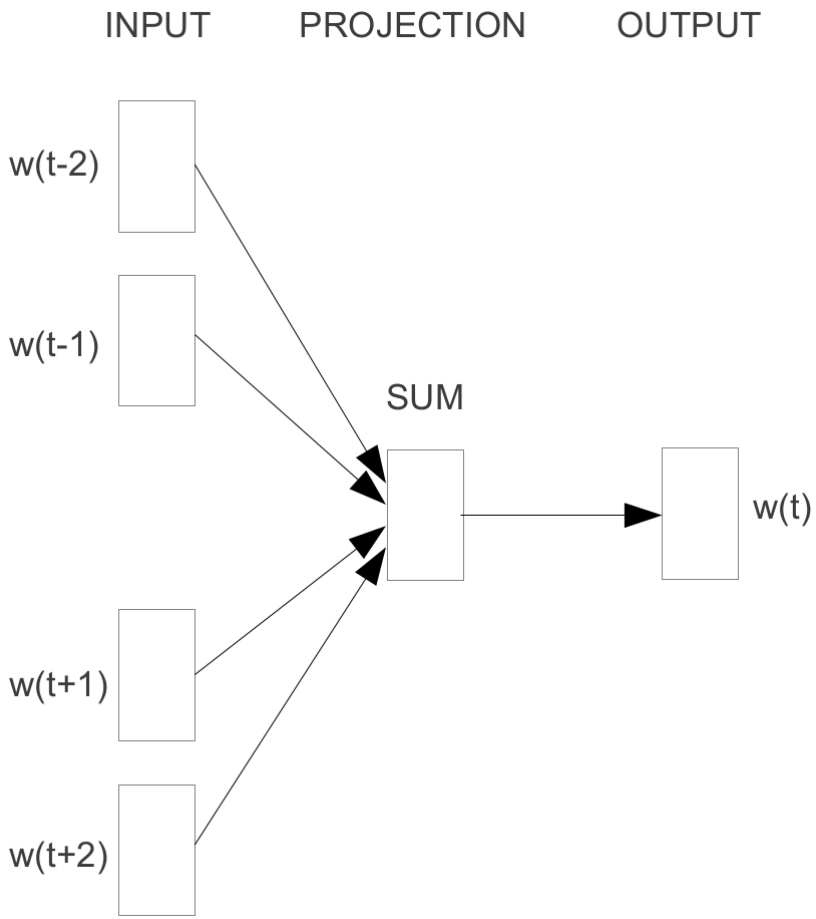
# Embedding matrix

Матрица представлений:

$$E = \begin{bmatrix} —e_1— \\ —e_2— \\ \cdots\cdots\cdots \\ —e_{|V|}— \end{bmatrix}$$

Каждая строка — представление одного слова.
Идея: обучим матрицу $E$ при помощи нейронной сети.

# Continuous bag of words



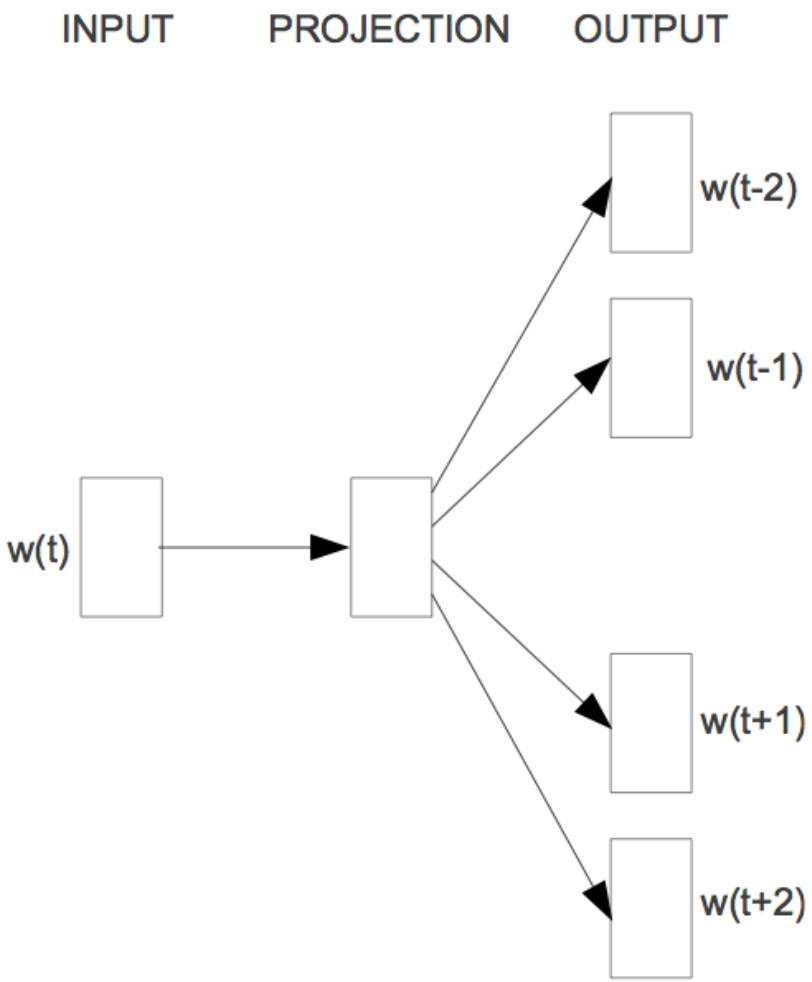

Предсказываем пропущенное слово по контексту:

$$P(t_i|\text{context}(w_i)) = \text{softmax}\left(\sum_{t_j \in \text{context}(w_i)} E_i W_v\right)$$

Функция потерь:

$$L = -\log P(w_i|\text{context}(w_i))$$

# Skip-gram




Предсказываем пропущенное слово по контексту:

$P(t_j|t_i) = \text{softmax}(E_i W_v)$

Функция потерь:

$$L = -\log P(\text{context}(t_i)|t_i) = -\sum_{t_j \in \text{context}(t_i)} logP(t_j|t_i)$$

# Проблема Softmax

- Linear + softmax: $P(t_i|X) = \frac{e^{h_i^T x_i}}{\sum_j e^{h_j^T x_j}}$
- Для вычисления градиента лог-лосса нужна только одна компонента $P(t_i|X)$
- Для вычисления одной компоненты надо вычислить все другие

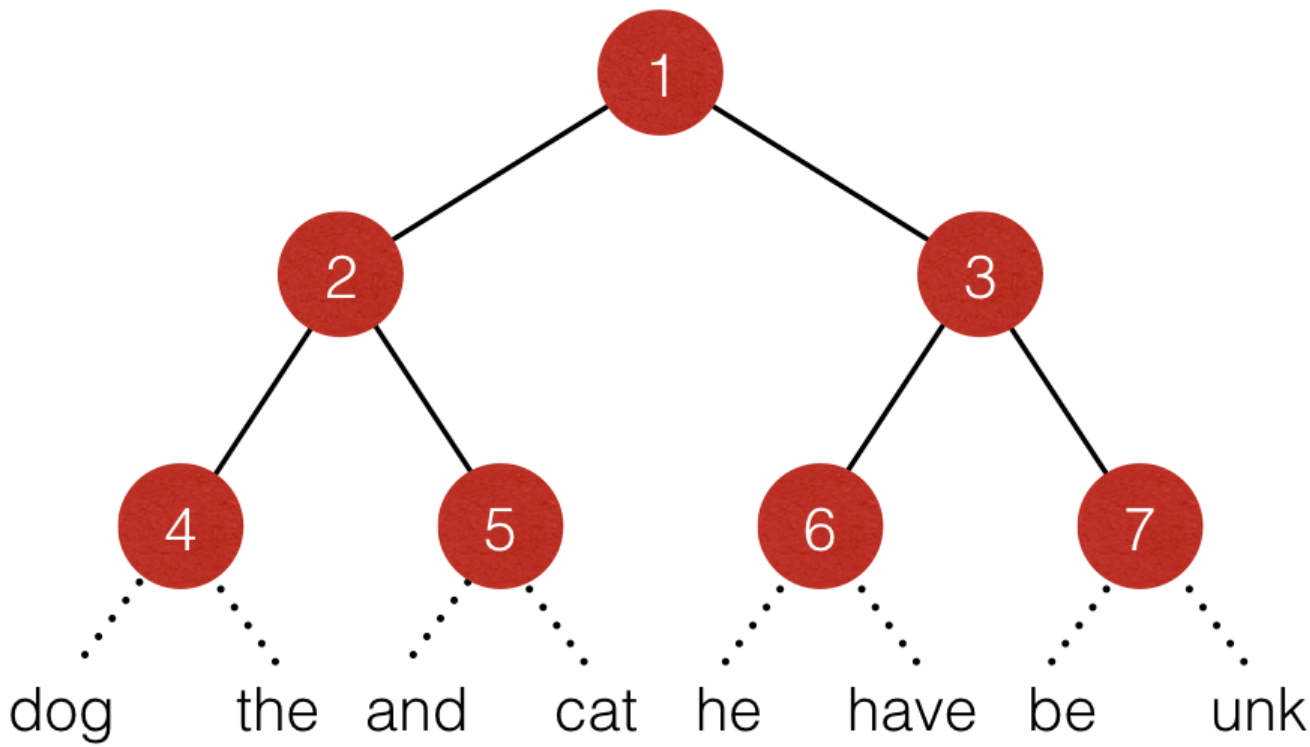
$\Rightarrow$ Медленная работа для больших словарей

Решения:

- Иерархический Softmax
- Дифференцированный Softmax
- Sampled Softmax / Noise-contrastive estimation

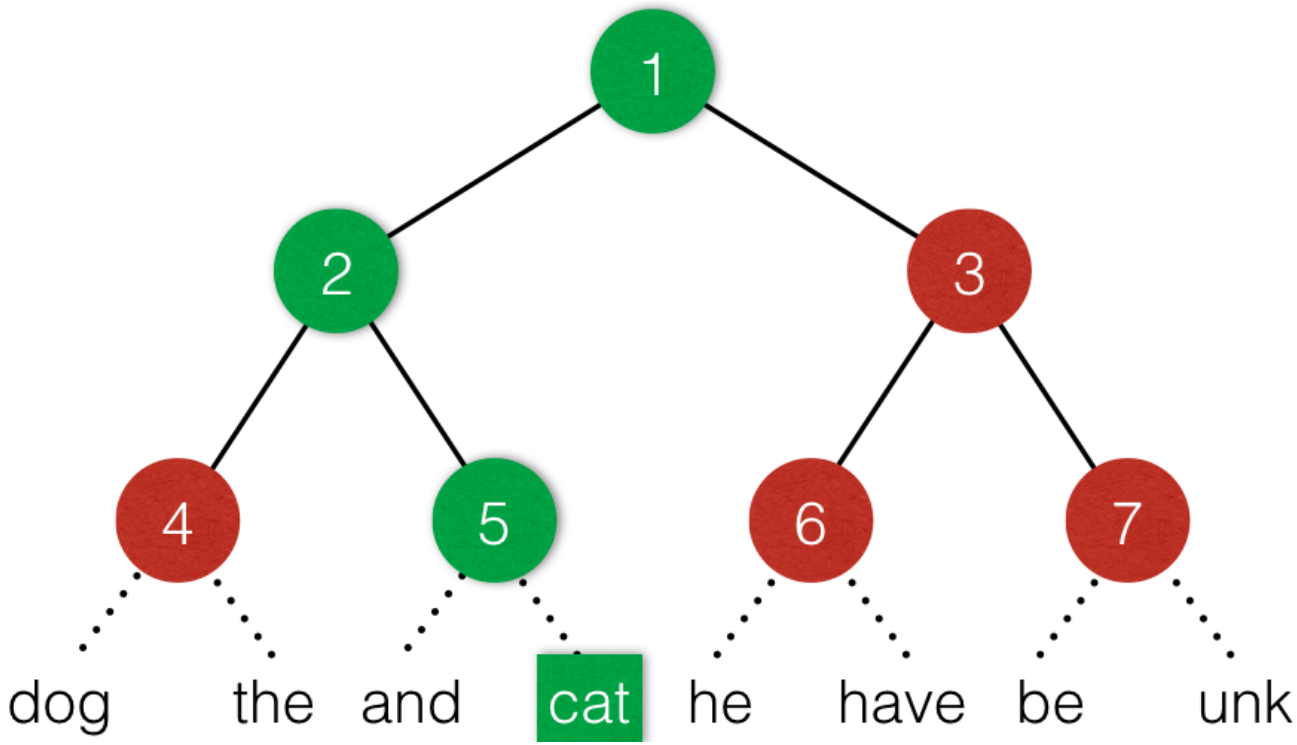
# Иерархический Softmax

$$P(\text{«cat»}|\text{context}) = P(1 \to 2|\text{context}) \times P(2 \to 5|\text{context}) \times P(5 \to \text{«cat»}|\text{context})$$



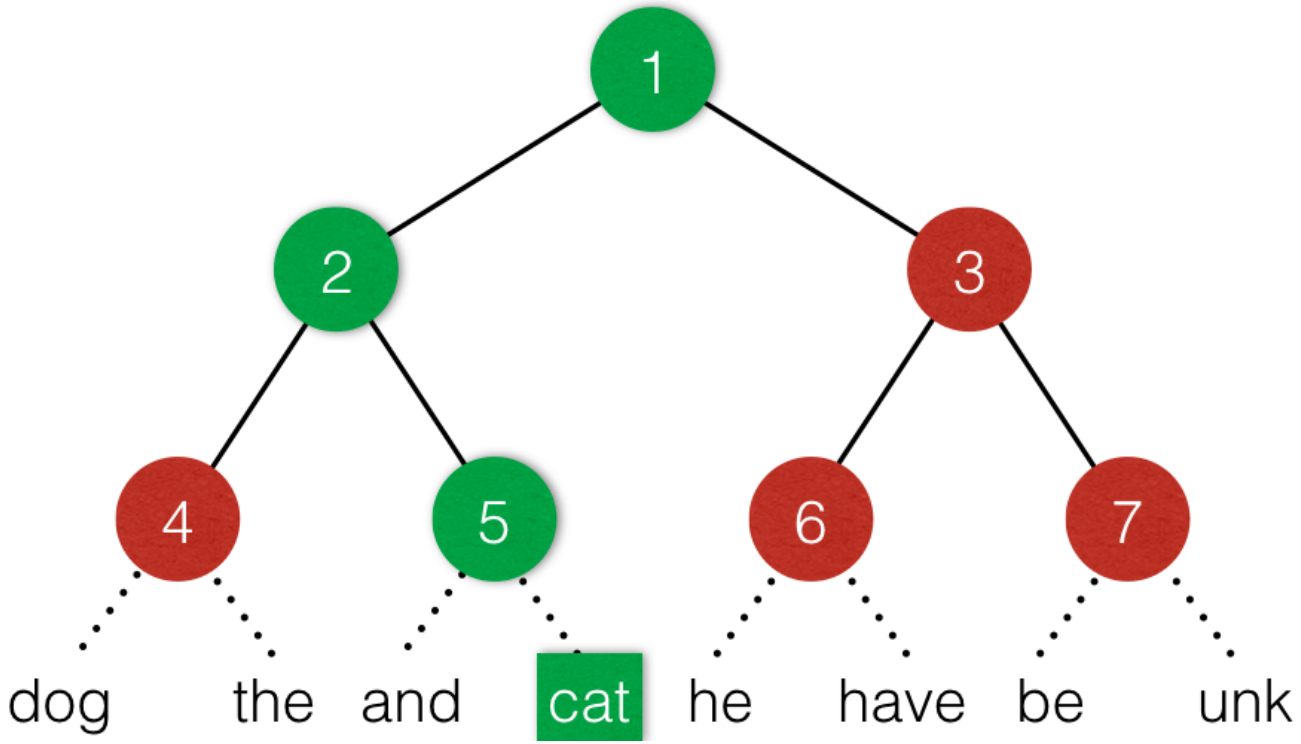
# Иерархический Softmax

$$P(\text{«cat»}|\text{context}) = P(1 \to 2|\text{context}) \times P(2 \to 5|\text{context}) \times P(5 \to \text{«cat»}|\text{context})$$



# Иерархический Softmax

$$P(\text{«cat»}|\text{context}) = (1 - \sigma(b_1 + h_1^T f(x))) \times \sigma(b_2 + h_2^T f(x)) \times \sigma(b_5 + h_5^T f(x))$$



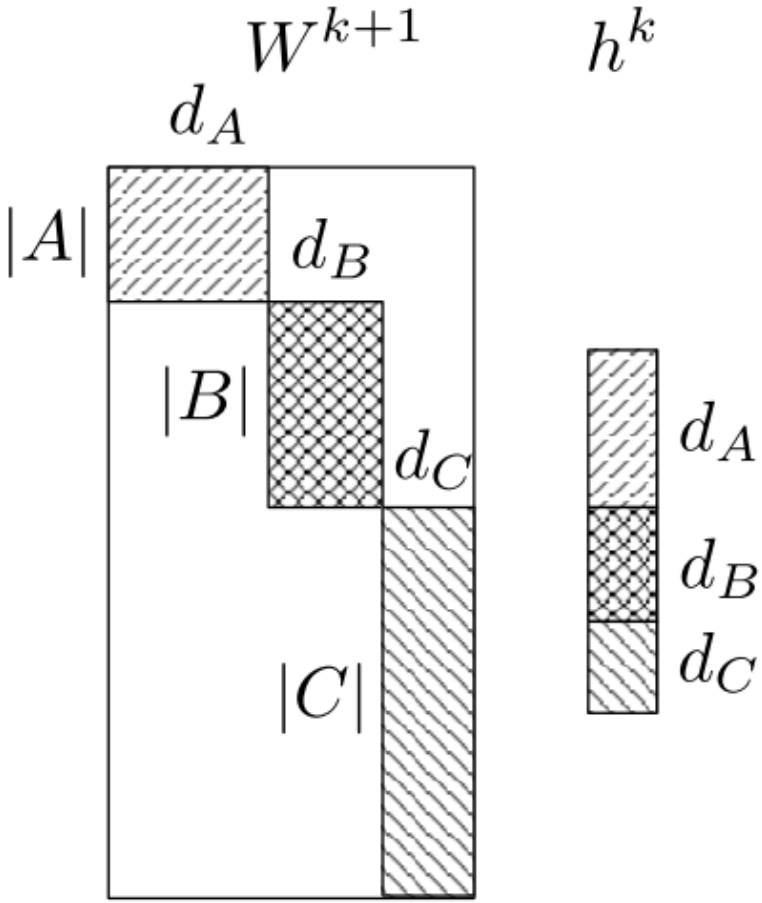
# Иерархический Softmax

Как построить дерево?

- Случайно приписать слова листьям
- Иерархическая кластеризация представлений слов
- Код Хаффмана

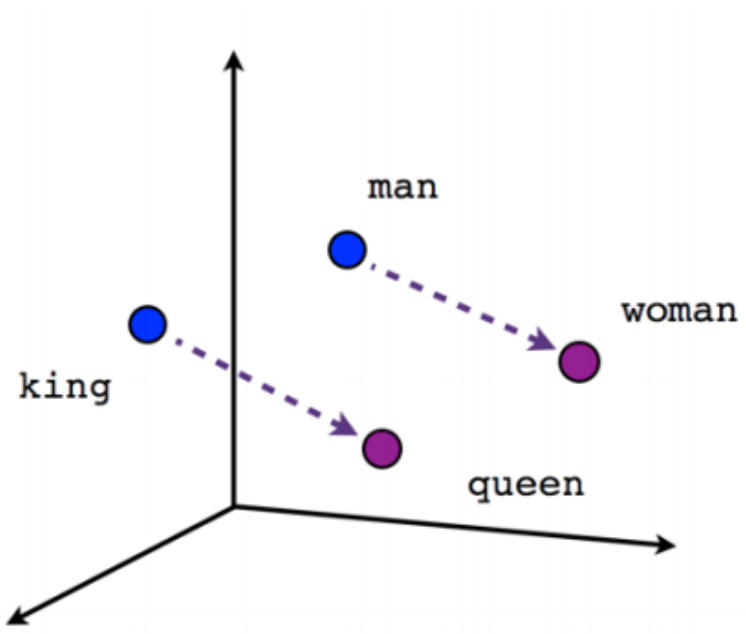
В train-time сложность $O(log_2 N)$
В test-time требуется вычислять значения во всех листах $\Rightarrow$ медленнее чем обычный softmax.

# Дифференцированный Softmax

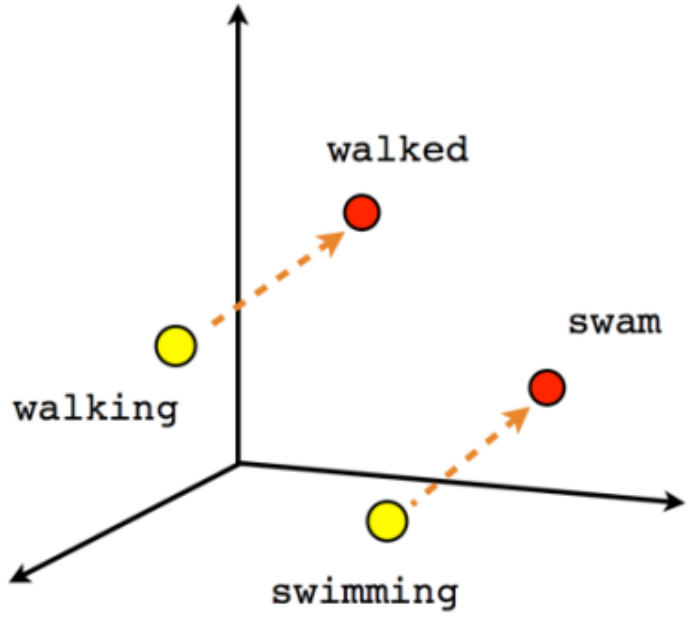



Редким словам сопоставляем короткие вектора, частым — длинные.
Скорость в train-time и test-time одинаковая.

# word2vec: арифметика




Male-Female




Verb tense

$e(\text{«king»}) - e(\text{«man»}) + e(\text{«woman»}) \simeq e(\text{«queen»})$

$e(\text{«swimming»}) + e(\text{«walked»}) - e(\text{«walking»}) \simeq e(\text{«swam»})$

# Чат-боты

# Известные боты

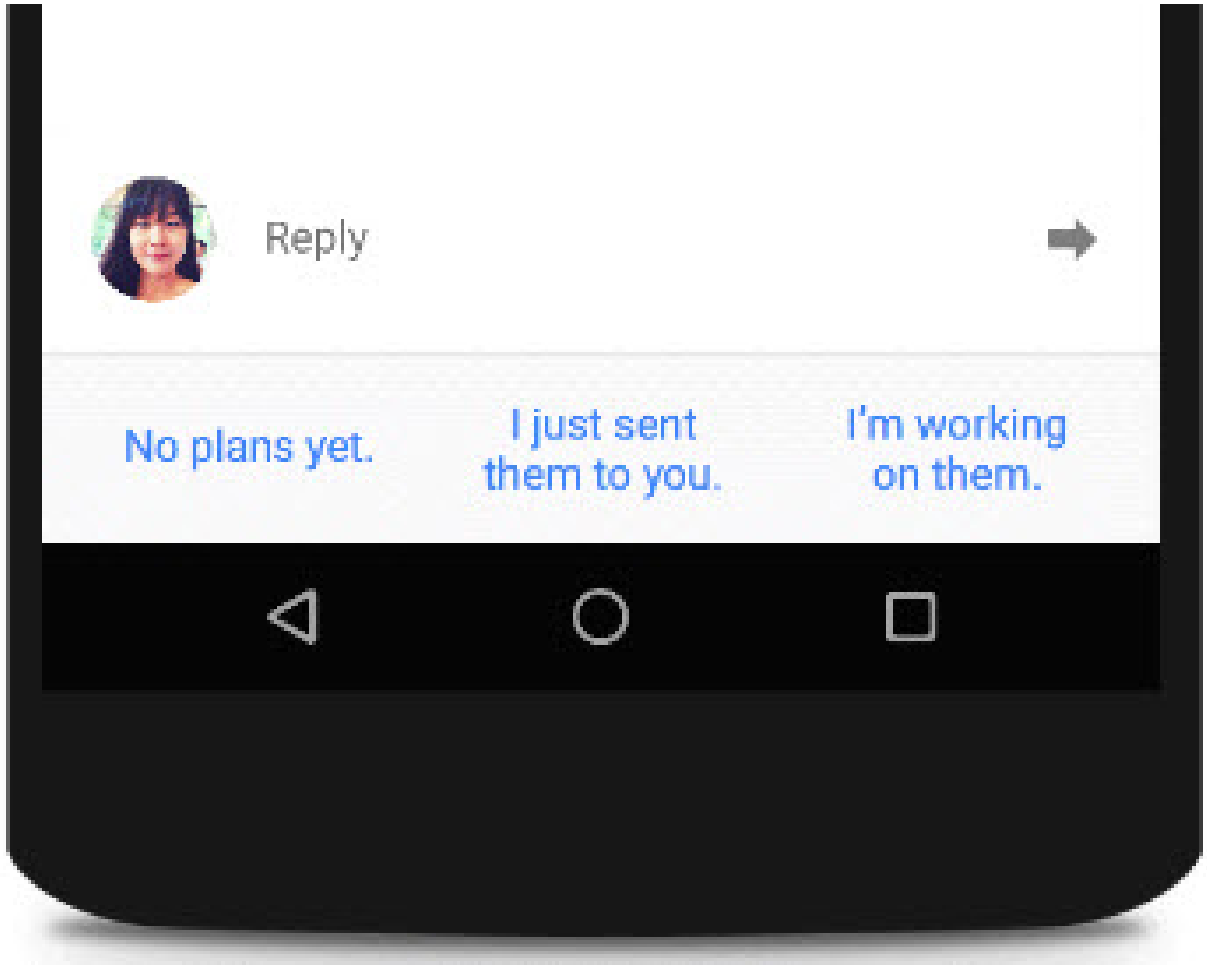



Рис.: Google smart reply

# Известные боты



Рис.: Microsoft Tay, твиттер бот

# Известные боты

В ответ 2 days til election

**DeepDrumpf** @DeepDrumpf · 31 окт.

I've got to win first. That's what I do. You have, right here in Colorado and a lot of the states, voting for ISIS on Nov 8th. @Donna_West

34 90

В ответ Jon Favreau

**DeepDrumpf** @DeepDrumpf · 26 окт.

No, Abraham Latino is poisoning our country. Impossible to sell our product. There will be no amnesty, but this is locker room talk.@jonfavs

57 127

Рис.: DeepDrumpf, твиттер бот

# Известные боты

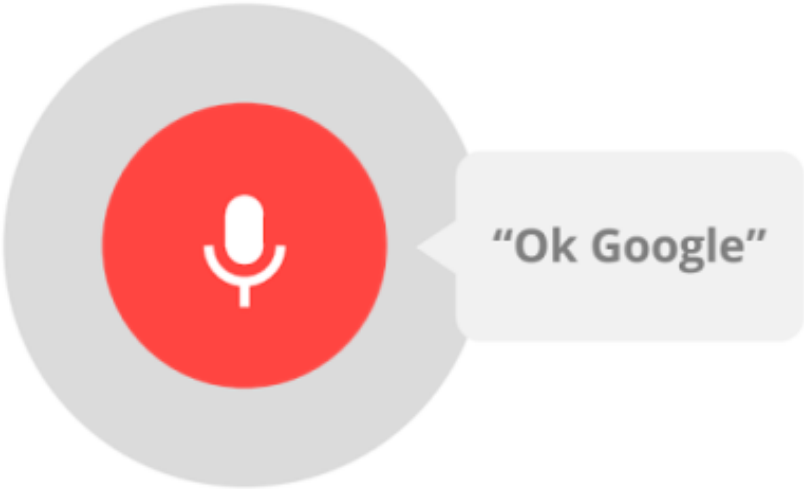
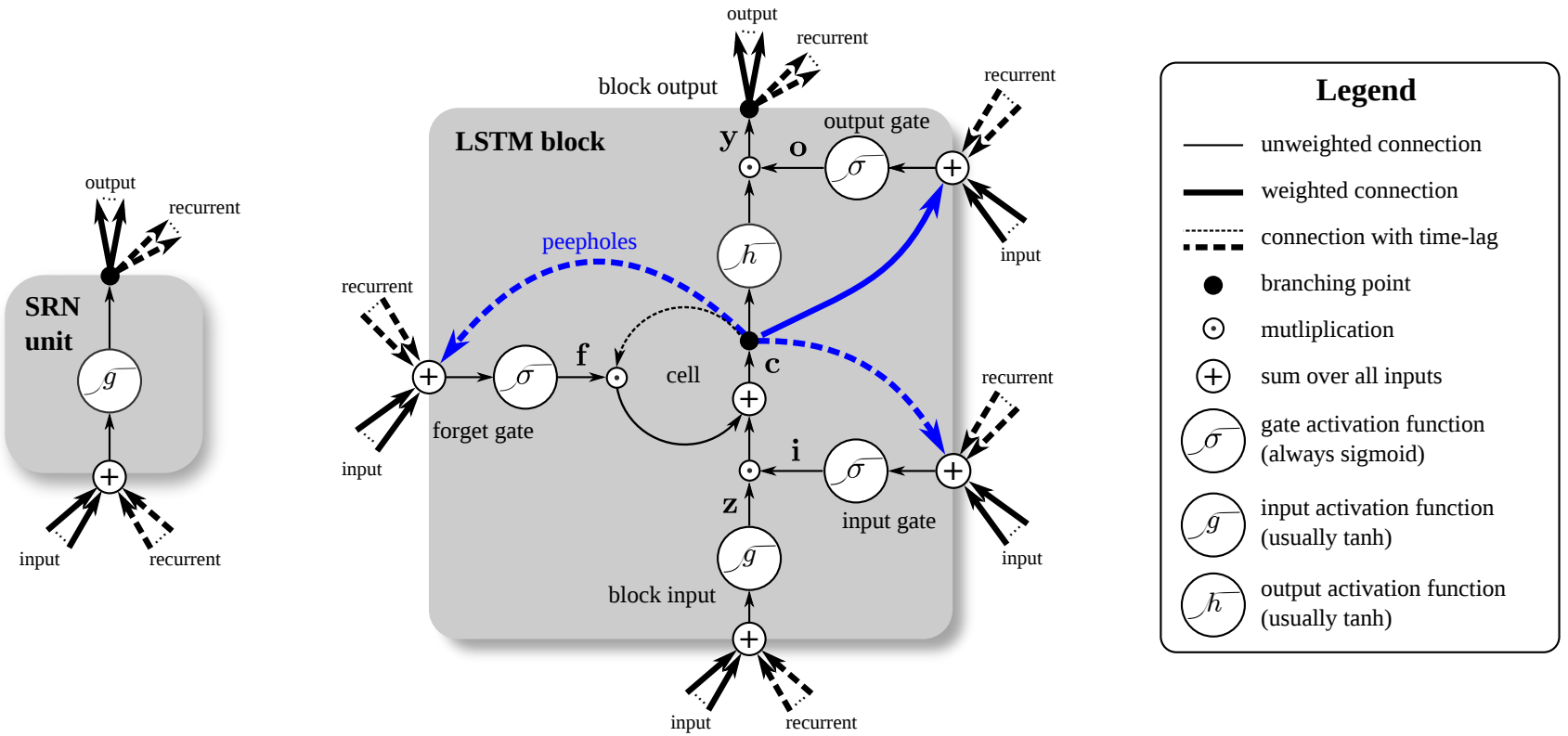



Рис.: Голосовые помощники

# LSTM сети


output
recurrent
SRN
unit
input
recurrent
output
recurrent
block output
LSTM block
output gate
recurrent
input
peepholes
recurrent
cell
forget gate
input
recurrent
input gate
input
block input
input
recurrent
Legend
unweighted connection
weighted connection
connection with time-lag
branching point
mutliplication
sum over all inputs
gate activation function (always sigmoid)
input activation function (usually tanh)
output activation function (usually tanh)

# Conversation vs Goal

## Общение для общения

- Ответы должны быть более-менее релевантными
- Надо поддерживать контекст беседы
- Ответы должны быть разнообразными
- Метрика: A/B тесты, асессоры

## Общение для достижения цели

- Реплики бота должны приближать диалог к цели
- Надо поддерживать контекст беседы
- Метрика: Accuracy, Precision, Recall, ...

# Вероятностная постановка

По последовательности слов $w_1, w_2, \ldots, w_n$ надо найти распределение $P(w_1, w_2, \ldots, w_n)$.

Chain rule: $P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, w_2, \ldots, w_{i-1})$

Требуется научиться генерировать следующее слово по предыдущим.

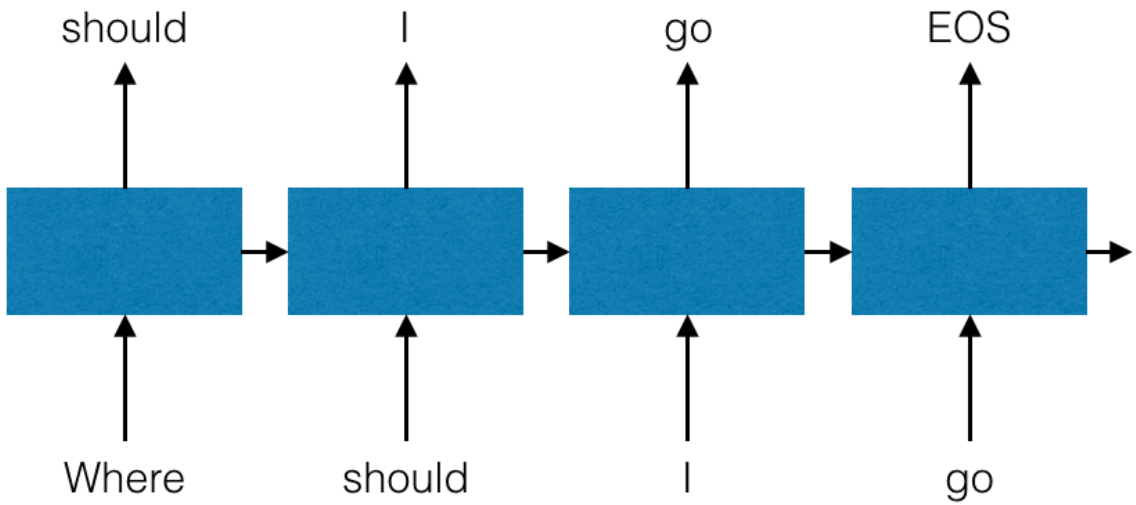
# N-grams

$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, \ldots, w_{i-1}) \simeq \prod_{i=1}^{n} P(w_i | w_{i-k}, \ldots, w_{i-1})$$

Обучение: считаем количество вхождений $w_{i-k}, \ldots, w_{i-1}, w_i$ и нормируем, чтобы получить вероятности.

- Чем больше k, тем более общая/переобученная модель
- Требуется много памяти
- Ограниченная длина контекста

# Нейронные сети




$$P(w_1, w_2, \ldots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, \ldots, w_{i-1})$$

$$P(w_i | w_1, \ldots, w_{i-1}) \simeq f(w_i | c(w_1, \ldots, w_{i-1}))$$

Где $f, c$ — нейронные сети

# Задача генерации ответа

По последовательности слов $q_i$ надо найти распределение на последовательность $a_i$: $P(a|q) = ?$.

# Seq2Seq

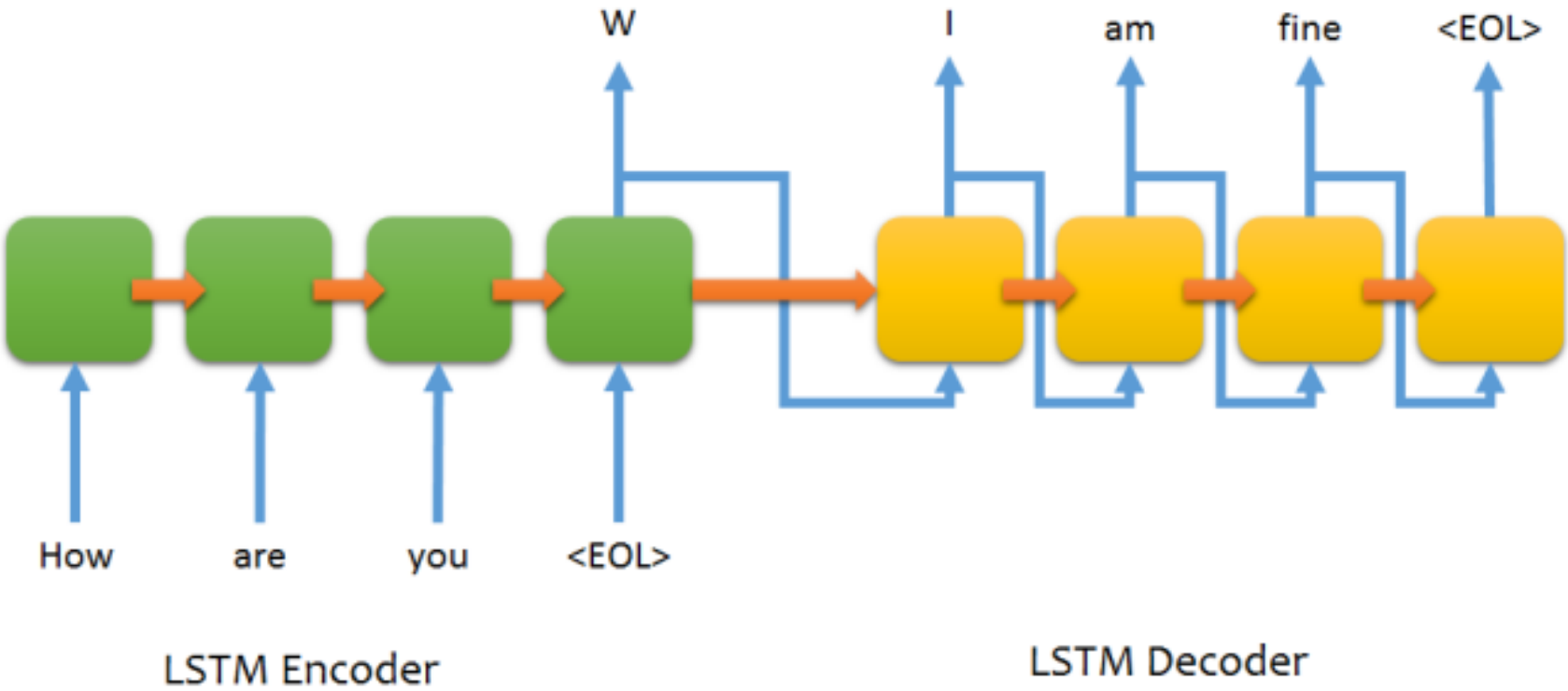



Рис.: Перевод последовательностей друг в друга

# Проблемы

## Генерация

Умеем вычислять $P(a|q)$ для всех возможных ответов.
Предсказание: $a_{MP} = \arg\max_a P(a|q)$ или $a_S \sim P(a|q)$.

Для $a_{MP}$ можно использовать beam search.

## Разнообразность

После beam search часто получаются частотные ответы: «да», «нет», «не знаю».
Можно обучить две сети: $P(a|q)$ и $P(q|a)$.

$$a_{MP} = \arg\max_a \Big[\lambda P(a|q) + (1-\lambda)P(q|a)\Big]$$

# За рамками лекции

- Поддержание диалога: HRED
- Attention
- Работа со словами не из словаря

# Вопросы