# Neural machine translation by jointly learning to align and translate

Свитанько Лиза

Факультет компьютерных наук,
Высшая школа экономики

Москва, 2017

# Машинный перевод

**Статистический машинный перевод (SMT)**: перевод генерируется на основе статистических моделей, параметры которых являются производными от анализа двуязычных корпусов текста.

Условная вероятность распределения последовательности $p(e|f)$,

$e$ - последовательность на языке, на который переводим (target),

$f$ - на языке для перевода (source).

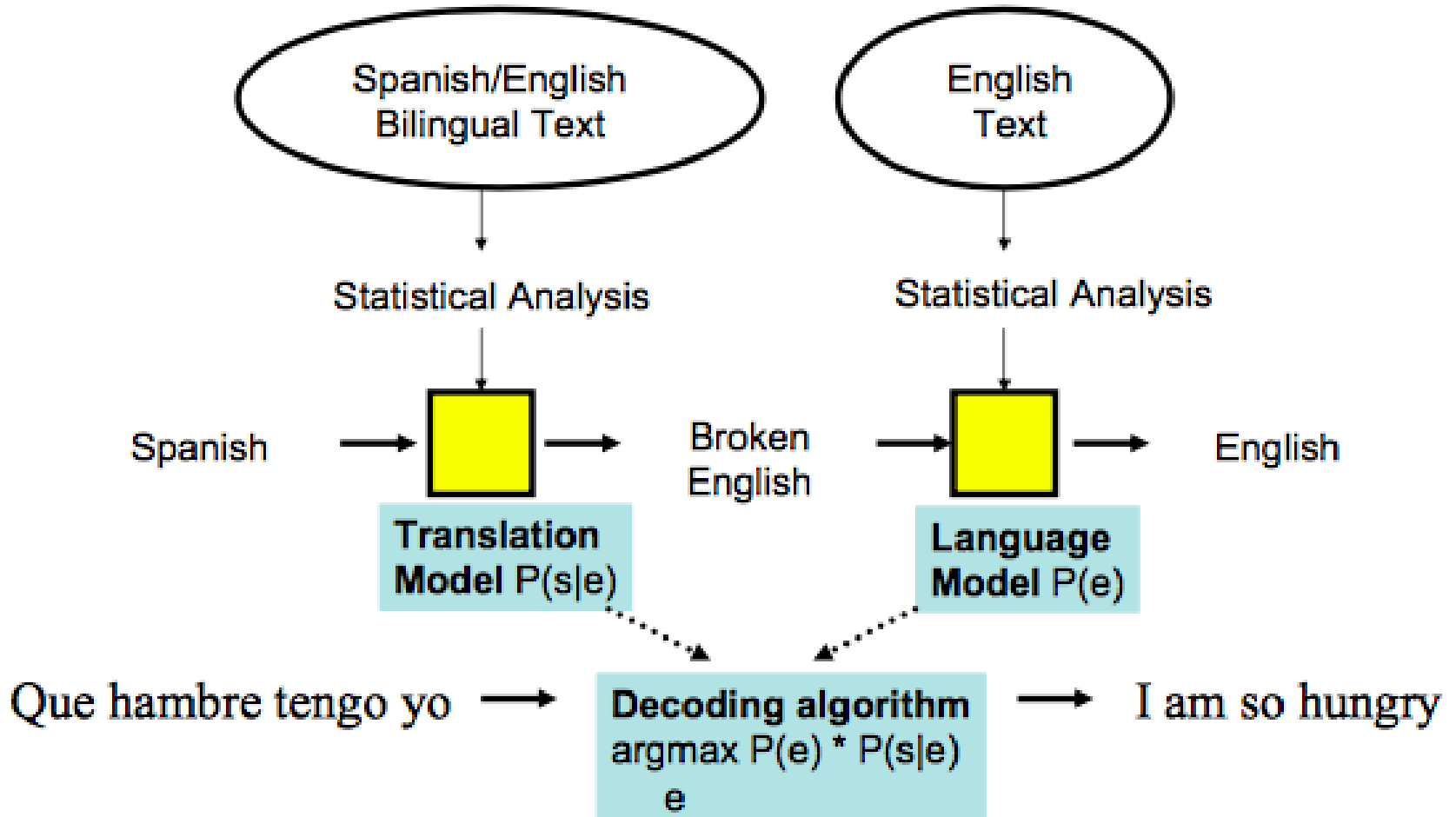
Теорема Байеса:

$$p(e|f) \propto p(f|e)p(e),$$

где $p(f|e)$ – вероятность f как перевода для e, $p(e)$ – вероятность встретить такую последовательность в виде таргета.
Задача:

$$\tilde{e} = arg \max_{e \in e^*} p(e|f) = arg \max_{e \in e^*} p(f|e)p(e)$$

- проходимся по всем последовательностям $e$ из языка $e^*$.


Spanish/English
Bilingual Text
English
Text
Statistical Analysis
Statistical Analysis
Spanish
Broken
English
English
Translation
Model P(s|e)
Language
Model P(e)
Que hambre tengo yo
Decoding algorithm
argmax P(e) * P(s|e)
e
I am so hungry

1. Модель для перевода (предпочтение более адекватным переводам)
   - *p(das haus ist klein|the house is small) >*
   - *p(das haus ist klein|the building is tiny) >*
   - *p(das haus ist klein|the shell is low)*
2. Языковая модель (логичное построение слов, например, n-gram)
   - *p(the house is small) > p(small the is house)*

# Word based SMT

- найти переводы в словаре
- посчитать количество употреблений в parallel corpus

| Translation of *Haus* | Count |
|---|---|
| house | 8,000 |
| building | 1,600 |
| home | 200 |
| household | 150 |
| shell | 50 |

- максимизация правдоподобия

$$p_f(e) = \begin{cases} 0.8 & \text{if } e = \text{house}, \\ 0.16 & \text{if } e = \text{building}, \\ 0.02 & \text{if } e = \text{home}, \\ 0.015 & \text{if } e = \text{household}, \\ 0.005 & \text{if } e = \text{shell}. \end{cases}$$

# SMT выводы

Плюсы:

- лучшие результаты благодаря использованию языковой модели
- в общем случае можно обобщать на разные пары языков
- много корпусов для обучения (bi-/mono-lingual)

Минусы:

- плохое качество на языках с разным порядком слов
- использование/создания большого корпуса
- нет возможности обобщения на языки разных языковых семейств

# Neural machine translation

Нейронный машинный перевод – машинный перевод с использованием нейросетей.

SMT + RNN Encoder-Decoder (Bi-RNN) в виде дополнительных признаков в существующую модель => модель учитывает синтаксические и семантические особенности текста и языка.

# RNN encoder-decoder

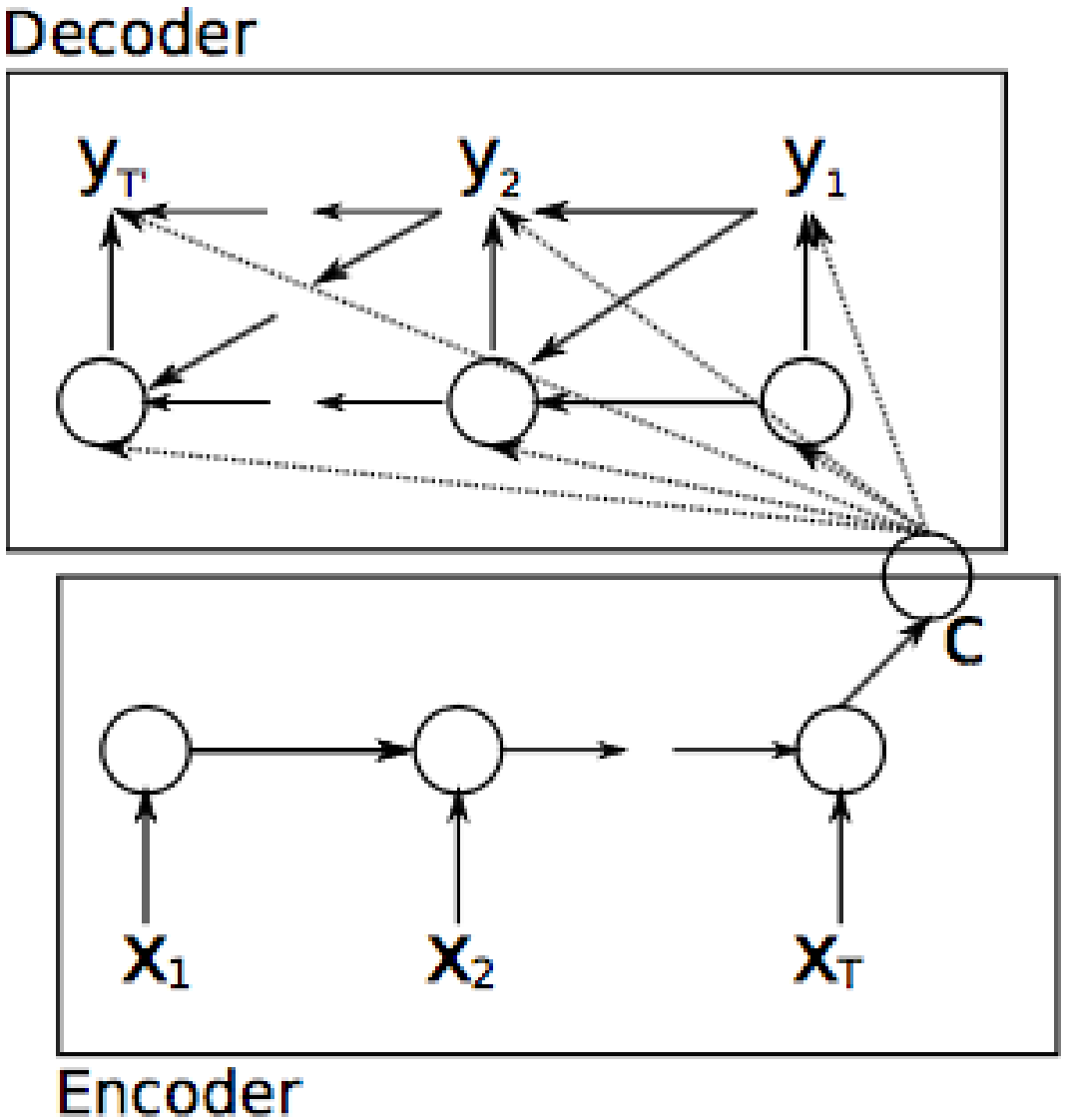
Условное распределение последовательности фиксированной длины: $p(y_1, \ldots, y_{T'} | x_1, \ldots, x_T)$, где T и T' могут отличаться.

1. encoder кодирует предложение в вектор фиксированной длины (скрытое состояние RNN на выходе)
2. decoder выдает перевод предложения из вектора

Скрытое состояние decoder в момент t:
$h(t) = f(h_{t-1}, y_{t-1}, c)$,
условное распределение следующего символа:

$$P(y_t | y_{t1}, y_{t2}, \ldots, y_1, c) = g(h_t, y_{t1}, c)$$

# RNN encoder-decoder



# RNN encoder-decoder

Обе компоненты совместно обучаются, максимизируя условное правдоподобие:

$$\max_{\theta} \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(y_n|x_n),$$

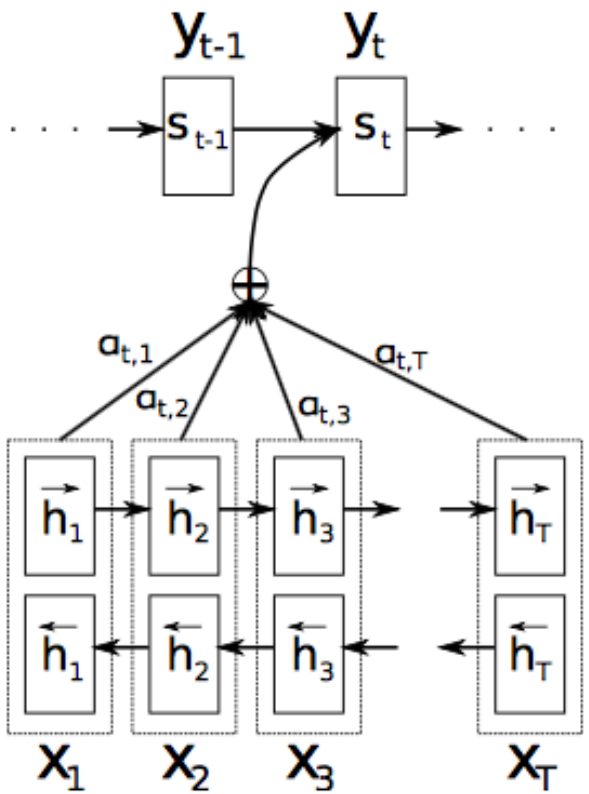
где $\theta$ – набор параметров, $(x_n, y_n)$ – пара (входящая посл-ть, выходящая посл-ть).

**Обученная модель:**

1. генерация перевода исходной последовательности
2. использование в оценивании условной вероятности $p_\theta(y|x)$ для SMT.

# Модификация

- Encoder - BiRNN
- Decoder - поиск по исходному предложению во время перевода



# Новый decoder

$$p(y_i|y_1, \ldots, y_{i-1}, x) = g(y_{i-1}, s_i, c_i),$$

где $s_i$ - скрытое состояние RNN в момент $i$ и

$$s_i = f(s_{i-1}, y_{i-1}, c_i)$$

$c_i$ - контекстный вектор для конкретного слова в переводе $y_i$.
$(h_1, \ldots, h_{T_x})$ - вектор аннотаций по каждому слову из исходного предложения.

# Контекстный вектор

Каждый $c_i$ как взвешенная сумма по аннотациям для $i$-го слова перевода:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j$$

Веса (матожидание аннотации):

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})}$$

где $e_{ij} = a(s_{i-1}, h_j)$ - "энергия похожести" $j$-го слова исходного предложения и $i$-го слова перевода (soft alignment).

# Новый encoder

Учет предыдущих слов и последующих слов (контекст) => bidirectional RNN.

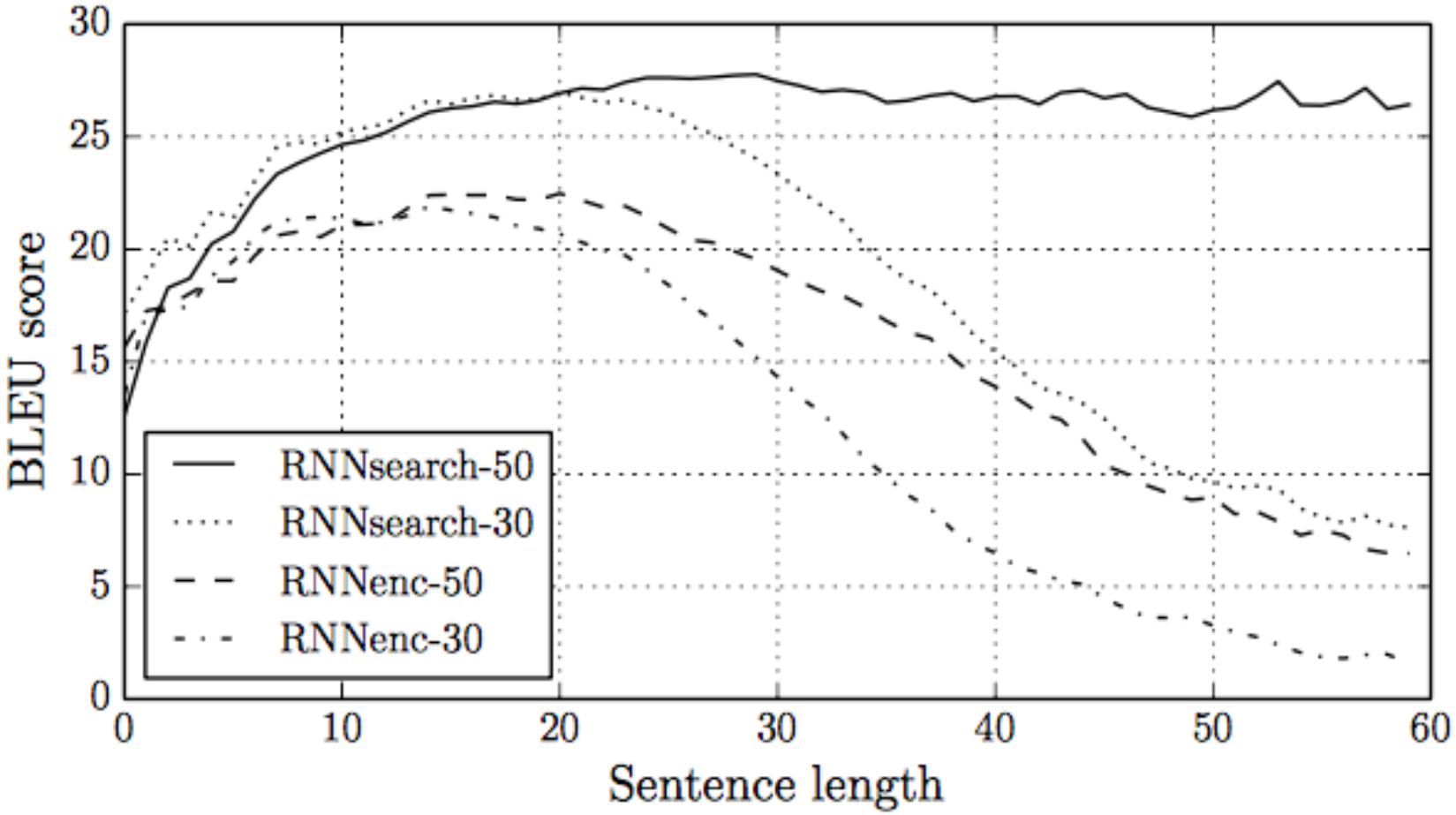
1. forward RNN: подсчет направленных вперед скрытых состояний $(h'_1, \ldots, h'_{T_x})$
2. backward RNN: подсчет направленных в обратную сторону скрытых состояний $(h''_1, \ldots, h''_{T_x})$
3. $h_j = [h'^T_j, h''^T_j]$ - финальная аннотация для $x_j$ с близко расположенным контекстом

# Эксперименты

English-to-French translation task, bilingual parallel corpora ACL WMT '14 — 348M слов.
Для сравнения две модели - encoder-decoder и search (описанная выше)

- Проверка на предложениях не длиннее 30 слов, затем не длиннее 50 слов
- 1000 скрытых юнитов в RNNencdec, 1000 скрытых юнитов на forward и backward RNN в encoder и еще 1000 в decoder для RNNsearch
- SGD + Adadelta для обучения

# Эксперименты



# Эксперименты

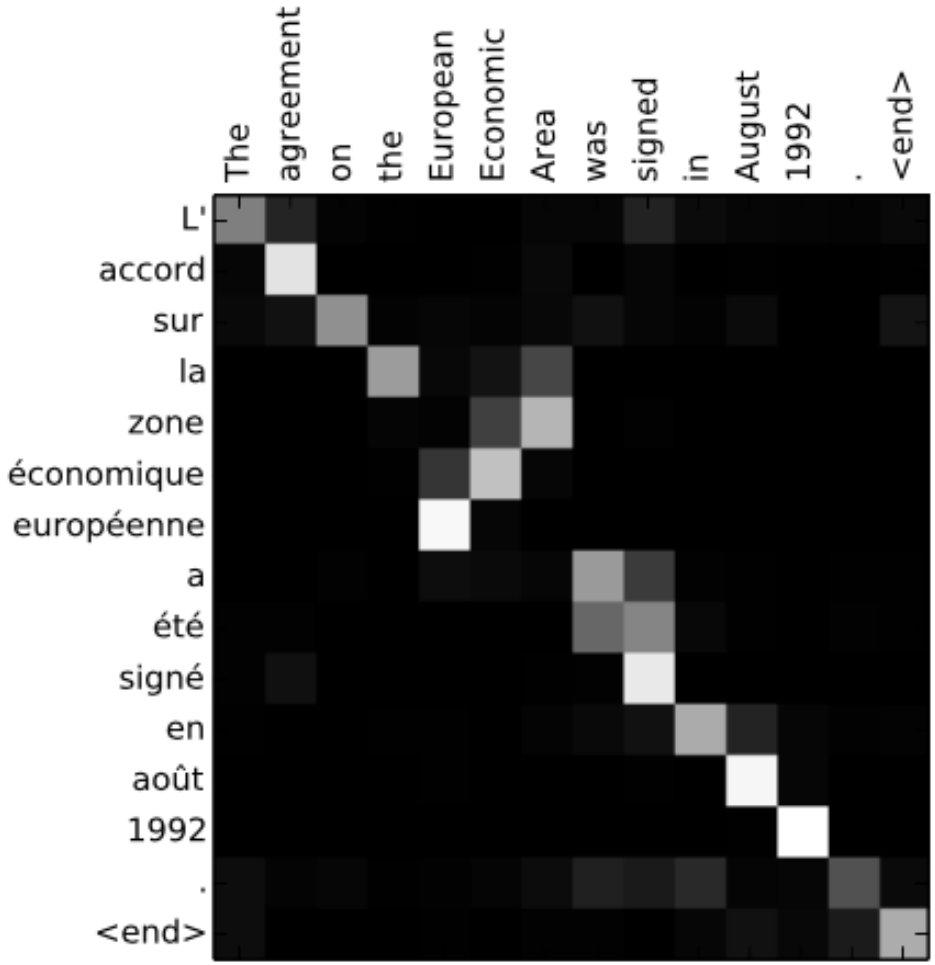



Рис.: Визуализация весов для каждой аннотации

# Заключение

- Для длинных предложений проблематично кодирование всего предлложения в один вектор фиксированной длины
- Модифицрованный encoder позволяет обращаться только к релевантной информации для каждого слова предсказания
- Модифицрованная модель сильно превышает по качеству базовый encoder-decoder, лучше работает на длинных предложениях и способна генерировать перевод с правильным расположением слов в зависимости от языка

Спасибо за внимание!