Машинный перевод

Аюпов Шамиль, 182

Задача

- S - source language, T - target language
- Предложения:

$$s = (s_1, \ldots, s_I),\ s_i \in S$$

$$t = (t_1, \ldots, t_J),\ t_i \in T$$

- Хотим осуществить “адекватный” перевод (сохранить смысл)

$$s \rightarrow t$$

Как оценивать перевод?

“Чем более похож машинный перевод на человеческий, тем лучше”

- Для каждого предложения есть профессиональный перевод (референс)
- Их может быть много, каждый со своим стилевым окрасом

- Идея оценивания - как-то оценить сходство перевода с референсным.

Как оценивать перевод?

BLEU (BiLingual Evaluation Understudy)

- Считается доля n-грамм, содержащихся хотя бы в одном референсе
- + Штраф за краткость (Brevity penalty)

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right). \qquad \text{BP} = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases} \qquad p_n = \frac{\sum_{\text{n-gram} \in \hat{y}} \text{count}_{\text{clip}}(\text{n-gram})}{\sum_{\text{n-gram} \in \hat{y}} \text{count}(\text{n-gram})}$$

TER (Translation Error Rate)

- Как сильно поменять перевод, чтобы получить референсный

$$\text{TER} = \frac{Nb\ (op)}{AvregN_{\ Ref}}$$

Where:

- Nb (op) : is the minimum number of edits;
- Avreg N_{ref}: the average size in words references.

Как оценивать перевод?

Проблемы:

- Плохо коррелируют с человеческим восприятием
- Не учитывают семантику и синтаксис предложения напрямую
- Не учитывают словоформы

Разные модификации пытаются исправить проблемы (METEOR, ROUGE, NIST, …)

Классические методы

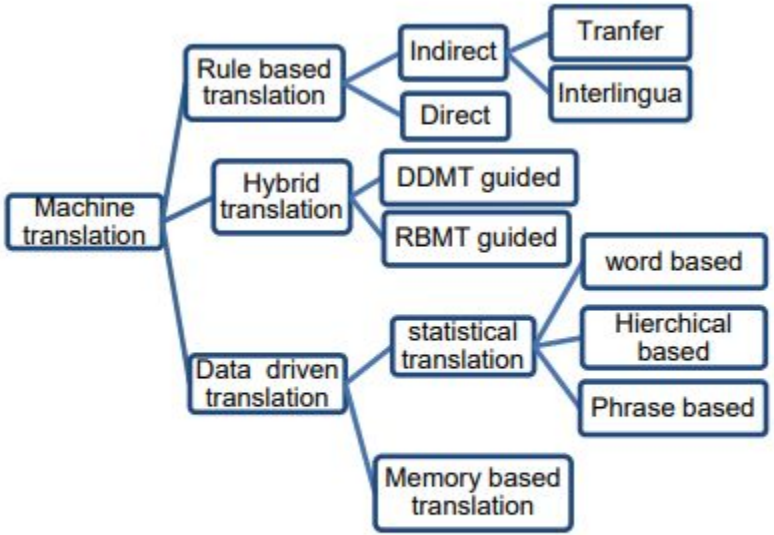
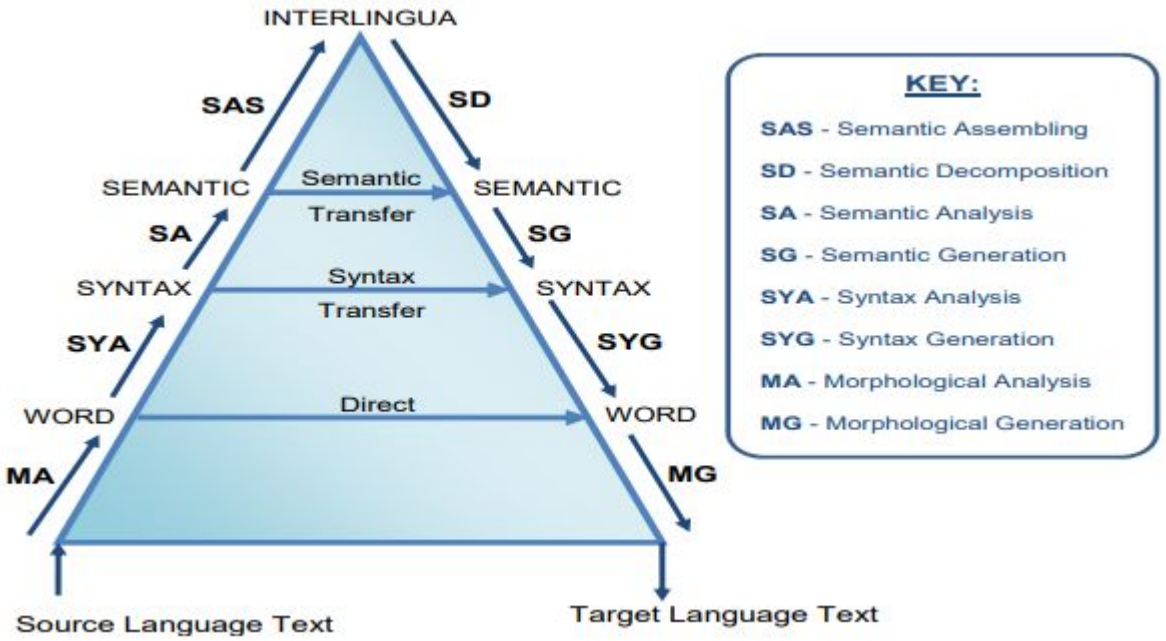


Figure 6 machine translation approaches

Rule-based approach



Особенности Rule-based approach

+ Не нужен параллельный корпус
+ Легко отлаживать
+ Не зависит от области (научные тексты и т.д.)

- Нужны специалисты
- Нужны хорошие словари
- Сложно расширять

Примеры: Apertium, GramTrans

Example-based MT

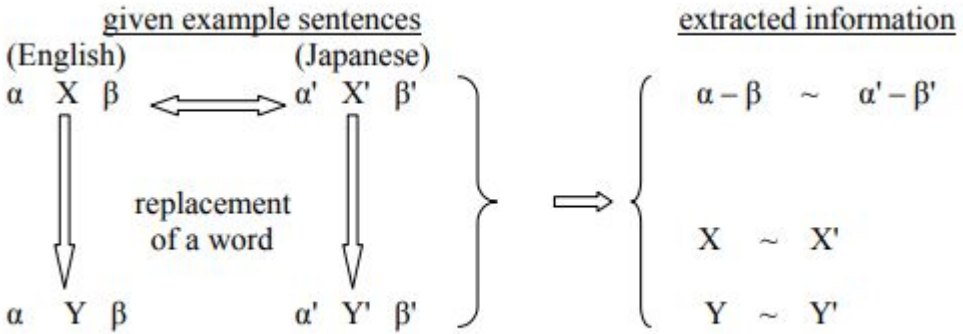


Fig. 1

Example of bilingual corpus

English	Japanese
How much is that **red umbrella**?	Ano **akai kasa** wa ikura desu ka.
How much is that **small camera**?	Ano **chiisai kamera** wa ikura desu ka.

Statistical MT

$$\text{argmax}_y P(y|x) = \underbrace{\text{argmax}_y}_{\textit{decoding}} \overbrace{P(x|y)}^{\text{Translation Model}} \overbrace{P(y)}^{\text{Language Model}}$$

Translation Model

Models how words and phrases should be translated (*fidelity*). Learnt from parallel data.

Language Model

Models how to write good English (*fluency*). Learnt from monolingual data.

Word based	Phrase based	Hierarchical phrases based

Особенности Data Driven approach

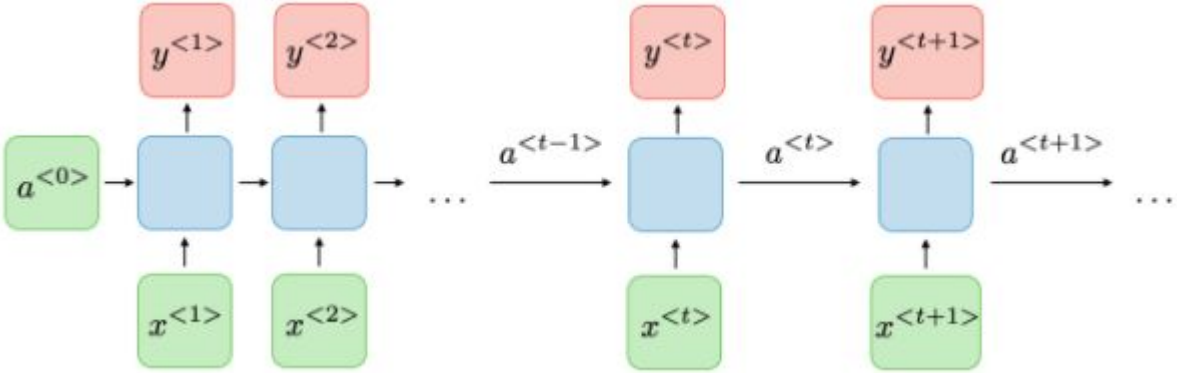
+ Не нужны специалисты для построения
+ Обычно не привязаны к отдельным языкам*
+ Конечное предложение более качественно из-за языковой модели

- Не всегда есть корпусы
- Сложно отлаживать
- Domain-dependent
- *Некоторые языки имеют особенности, требующие дополнительной работы

Примеры: GIZA++

RNN Recap

+ Последовательности любой длины
+ Shared параметры
+ Учет информации из прошлого
- Долгие вычисления
- Сложно учитывать информацию из далекого прошлого
- Не учитывает информацию о будущем



$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)$$

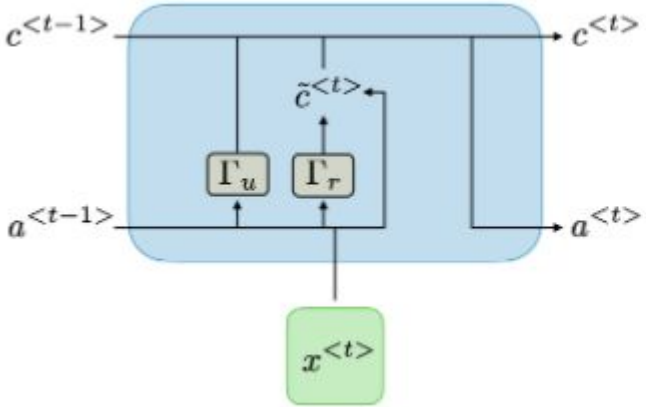
$$y^{<t>} = g_2(W_{ya}a^{<t>} + b_y)$$

$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}(\hat{y}^{<t>}, y^{<t>})$$

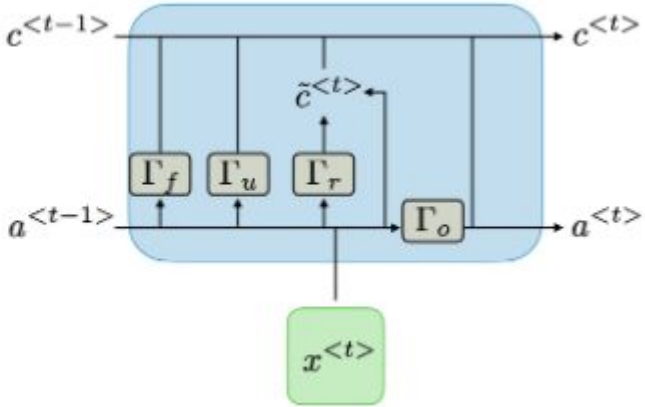
RNN Recap

Update gate Γ_u	How much past should matter now?
Relevance gate Γ_r	Drop previous information?
Forget gate Γ_f	Erase a cell or not?
Output gate Γ_o	How much to reveal of a cell?

$$\Gamma = \sigma(W x^{<t>} + U a^{<t-1>} + b)$$

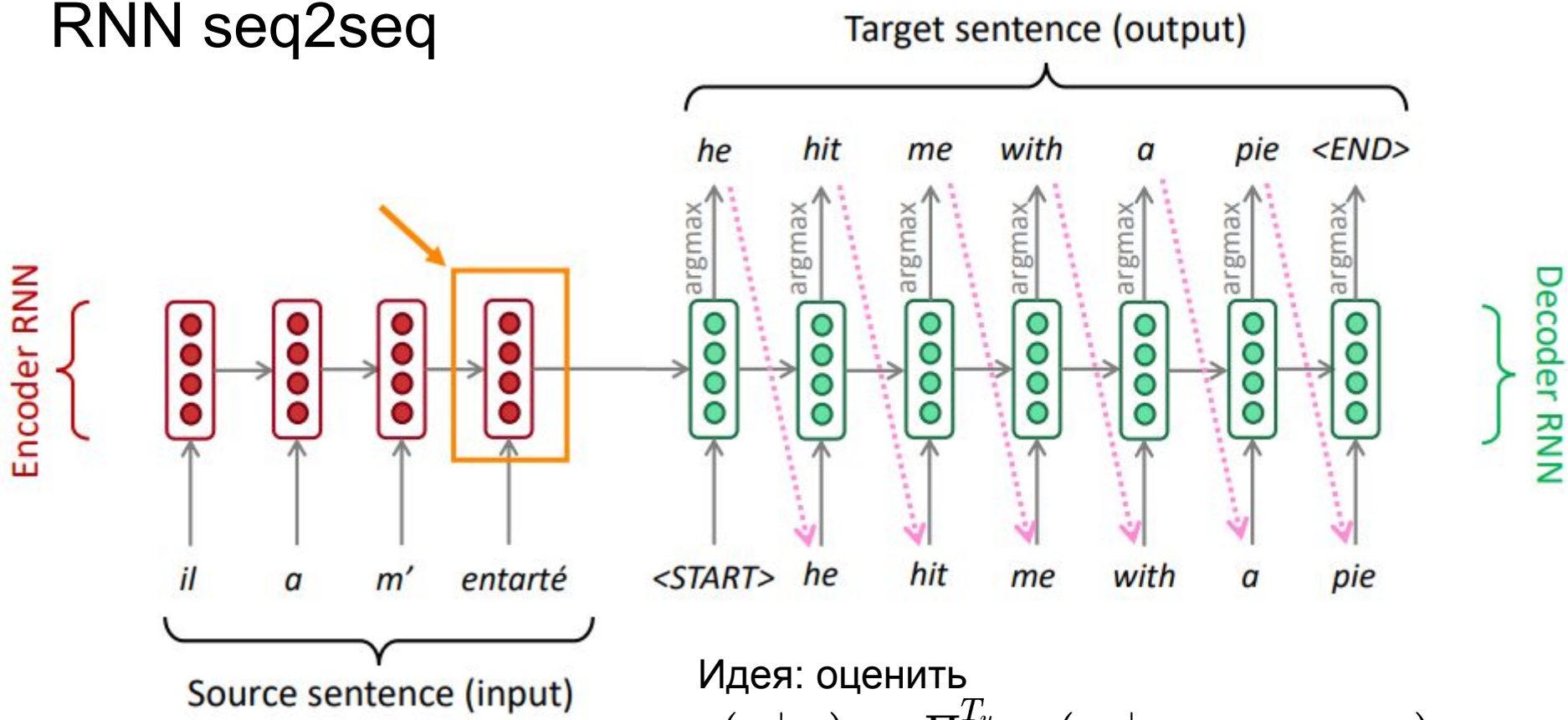


Gated Recurrent Unit



Long Short-Term Memory

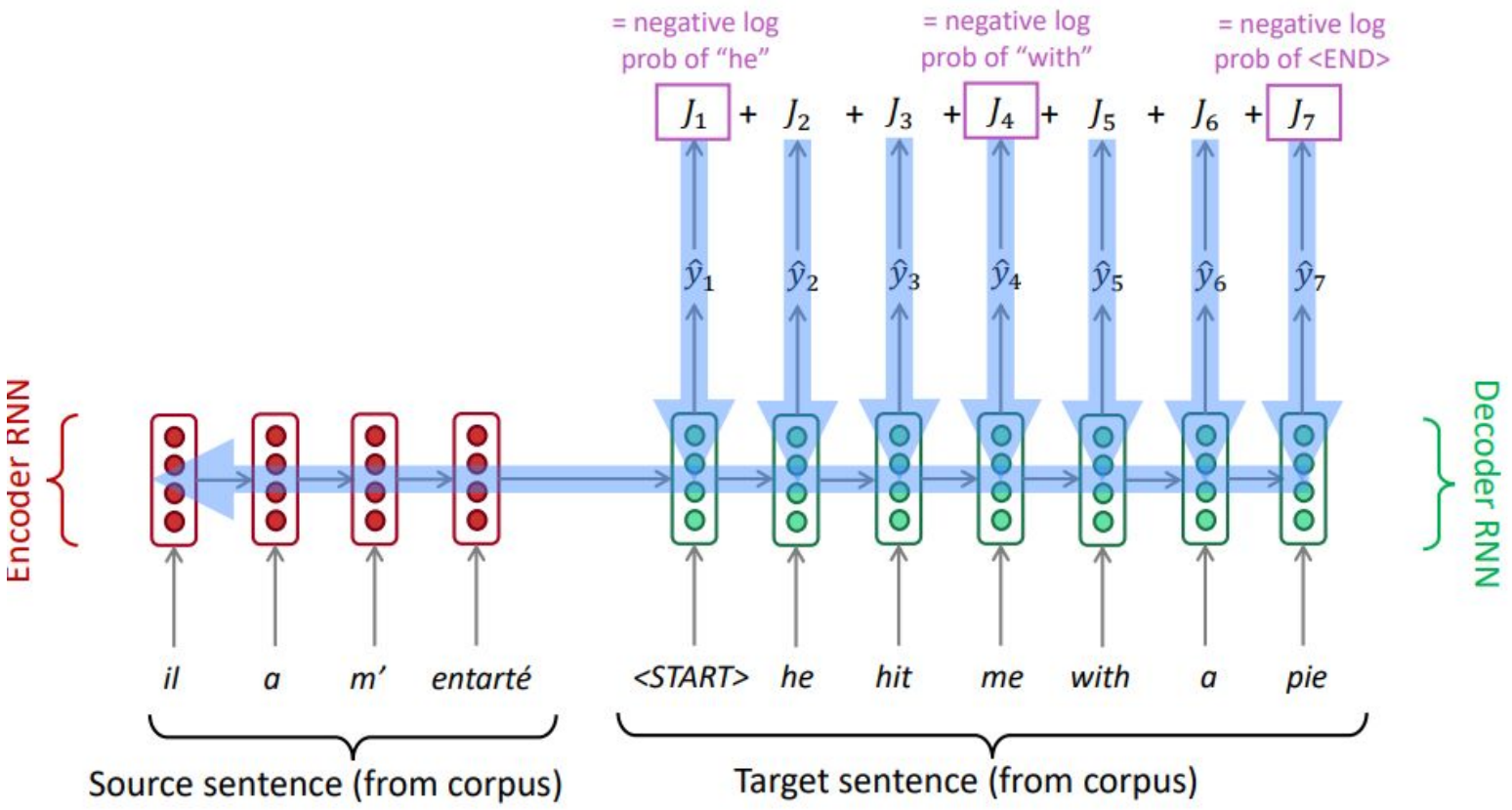
RNN seq2seq



Идея: оценить

$$p(y \mid x) = \Pi_{i=1}^{T_y} p(y_i \mid x, y_1 \ldots, y_{i-1})$$

RNN seq2seq: обучение

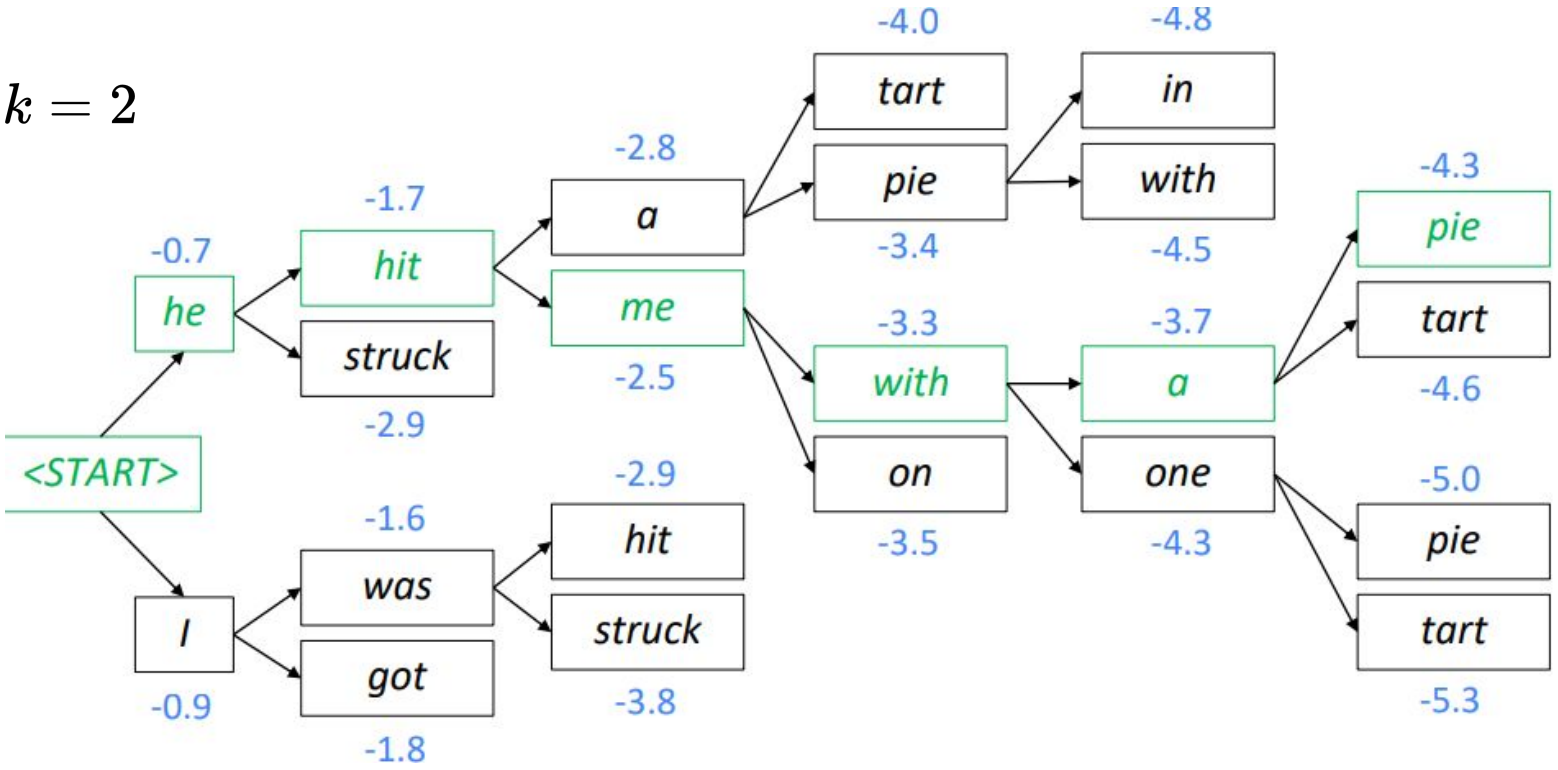


RNN seq2seq: предсказание

- Хотим найти $\arg\max_y p(y \mid x)$
- Искать жадно - плохо, один неверный выбор влечет последствия
- Выход - **beam-search**
  - Фиксируем k
  - Каждый кандидат длины t формирует top-k кандидатов длины t + 1
  - Среди k^2 кандидатов также выбираем top k
  - Если какое-то предложение закончилось, добавляем в пул гипотез
  - Продолжаем поиск пока в пуле не будет достаточно гипотез или не достигнем максимальной длины предложения
  - Отбираем наиболее вероятную из пула (с нормализацией)

Beam search: пример

$k = 2$



RNN seq2seq: summary

+ Достойное качество
+ Цельная система без отдельных компонент
+ Один подход для любой пары языков

- Нужен большой параллельный корпус
- Еще сложнее интерпретировать
- Часть проблем SMT (domain, out-of-vocabulary)
- *

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
State of the art [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	~45

RNN seq2seq: summary

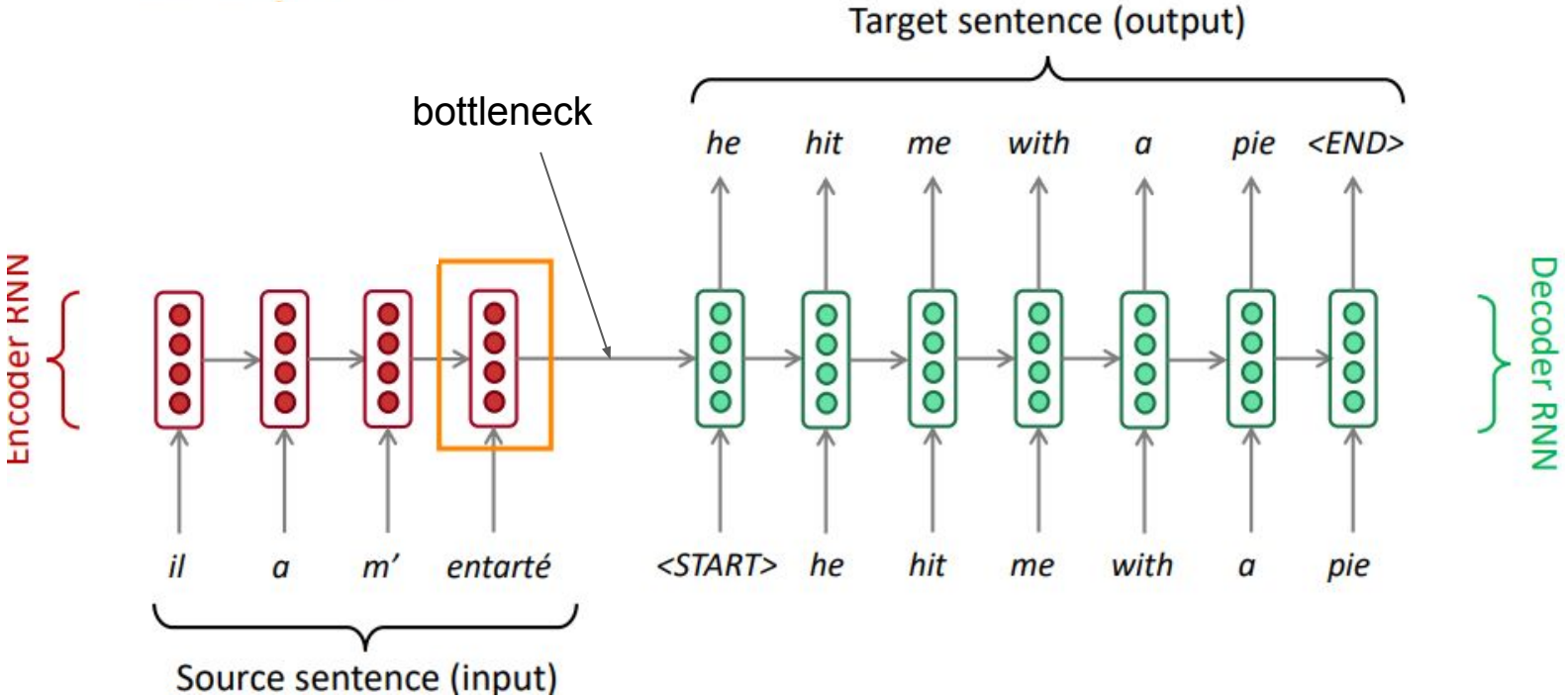
+ Достойное качество
+ Цельная система без отдельных компонент
+ Один подход для любой пары языков

- Нужен большой параллельный корпус
- Еще сложнее интерпретировать
- Часть проблем SMT (domain, out-of-vocabulary)
- Сложно переносить длинный контекст

Method	test BLEU score (ntst14)
Bahdanau et al. [2]	28.45
Baseline System [29]	33.30
Single forward LSTM, beam size 12	26.17
Single reversed LSTM, beam size 12	30.59
Ensemble of 5 reversed LSTMs, beam size 1	33.00
Ensemble of 2 reversed LSTMs, beam size 12	33.27
Ensemble of 5 reversed LSTMs, beam size 2	34.50
Ensemble of 5 reversed LSTMs, beam size 12	**34.81**

Method	test BLEU score (ntst14)
Baseline System [29]	33.30
Cho et al. [5]	34.54
State of the art [9]	**37.0**
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	**36.5**
Oracle Rescoring of the Baseline 1000-best lists	~45

Attention: мотивация



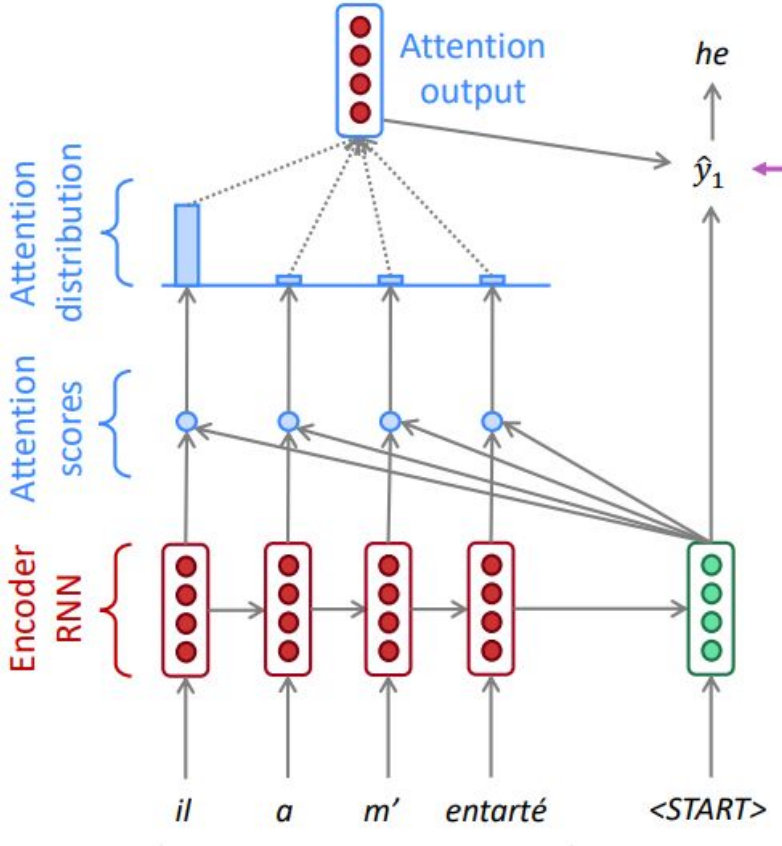
Идея: обеспечим прямой доступ ко всей информации

Attention

Интуитивно: способ выделить наиболее важную информацию из набора в зависимости от запроса

1. Используя N объектов и 1 запрос, считаем attention scores (N штук)
2. Считается attention distribution (softmax)
3. Берется взвешенная этим распределением сумма объектов -> attention output

Attention в RNN seq2seq



Concatenate attention output with decoder hidden state, then use to compute $\hat{y}_1$ as before

Attention score:

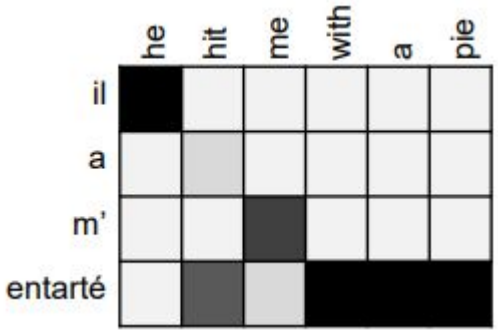
$$e_i = s^T h_i$$

$$e_i = s^T W h_i$$

$$e_i = v^T \tanh\left(W_1 h_i + W_2 s\right)$$

Attention: summary

- Улучшает качество
- Решает проблему bottleneck
- Интерпретируемость (alignment)
- Долго работает и плохо параллелится



System	Ppl	BLEU
Winning WMT'14 system – *phrase-based + large LM* (Buck et al., 2014)		20.7
Existing NMT systems		
RNNsearch (Jean et al., 2015)		16.5
RNNsearch + unk replace (Jean et al., 2015)		19.0
RNNsearch + unk replace + large vocab + *ensemble* 8 models (Jean et al., 2015)		**21.6**
Our NMT systems		
Base	10.6	11.3
Base + reverse	9.9	12.6 (*+1.3*)
Base + reverse + dropout	8.1	14.0 (*+1.4*)
Base + reverse + dropout + global attention (*location*)	7.3	16.8 (*+2.8*)
Base + reverse + dropout + global attention (*location*) + feed input	6.4	18.1 (*+1.3*)
Base + reverse + dropout + local-p attention (*general*) + feed input	5.9	19.0 (*+0.9*)
Base + reverse + dropout + local-p attention (*general*) + feed input + unk replace		20.9 (*+1.9*)
Ensemble 8 models + unk replace		**23.0** (*+2.1*)

Table 1: **WMT'14 English-German results** – shown are the perplexities (ppl) and the *tokenized* BLEU

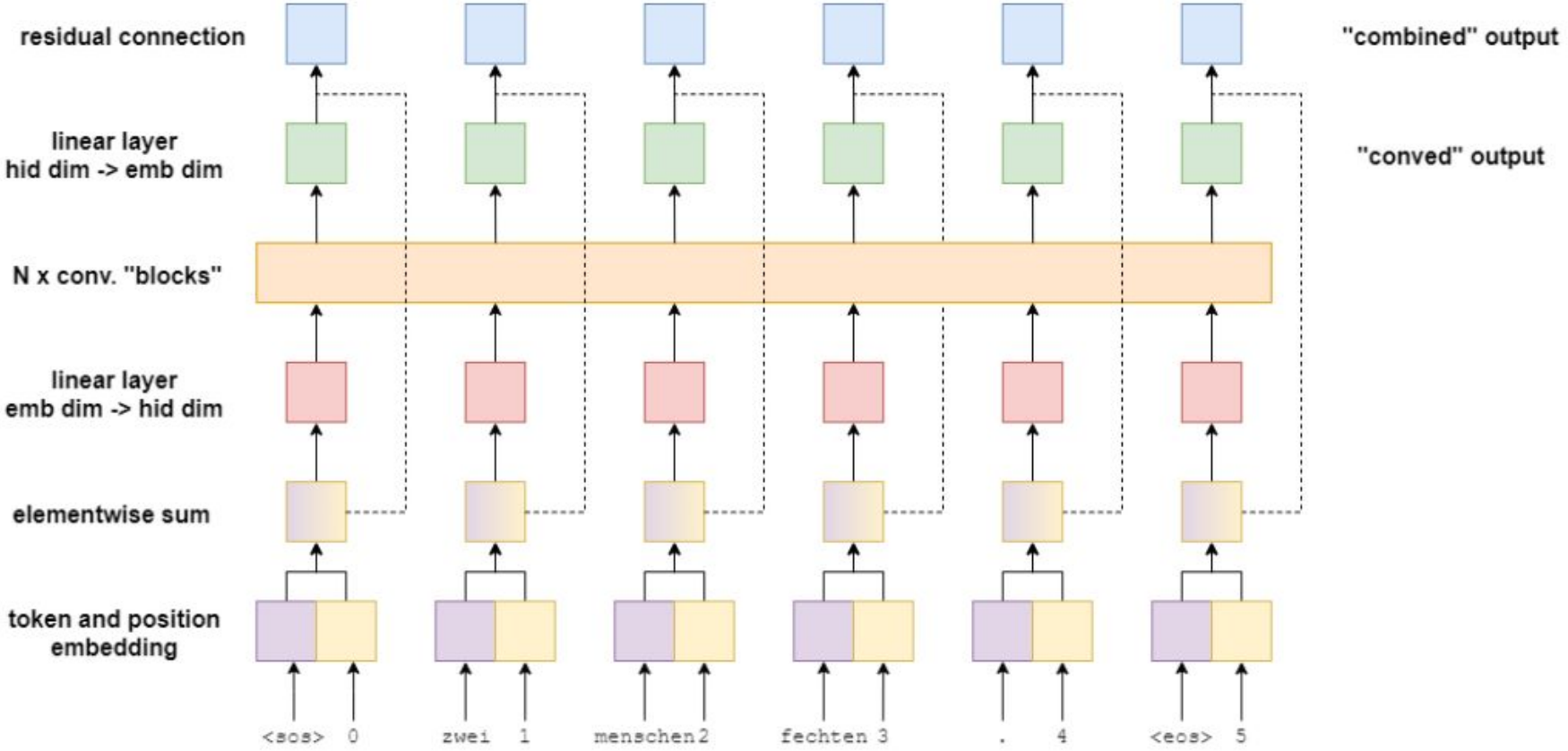
Conv seq2seq: начало

- Возможность распараллеливать - хорошо
- Хотим контролировать длину зависимости между словами
- Контролировать количество нелинейности тоже было бы хорошо

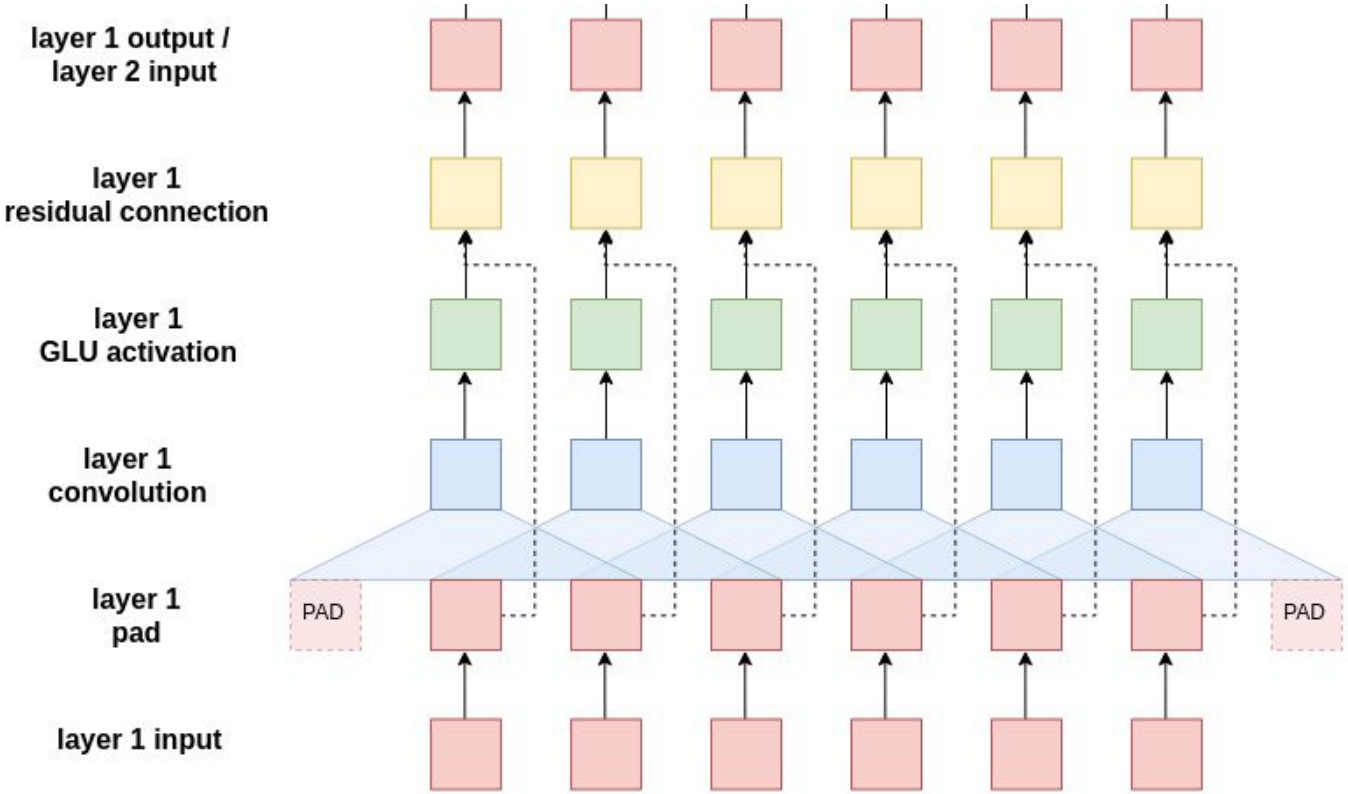
Gated Linear Unit (GLU):

$$GLU([A, B]) = A \bigotimes \sigma(B)$$

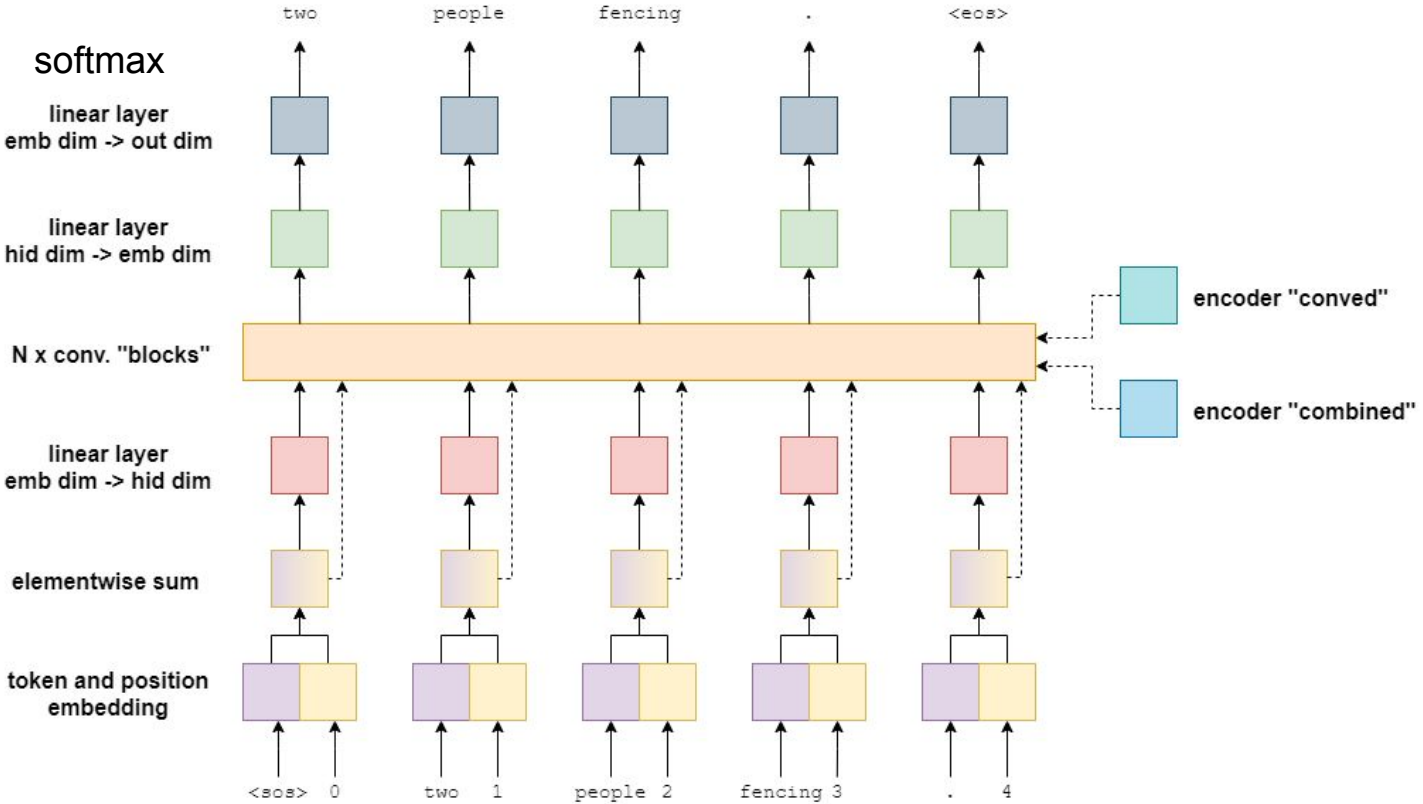
Conv seq2seq: Encoder



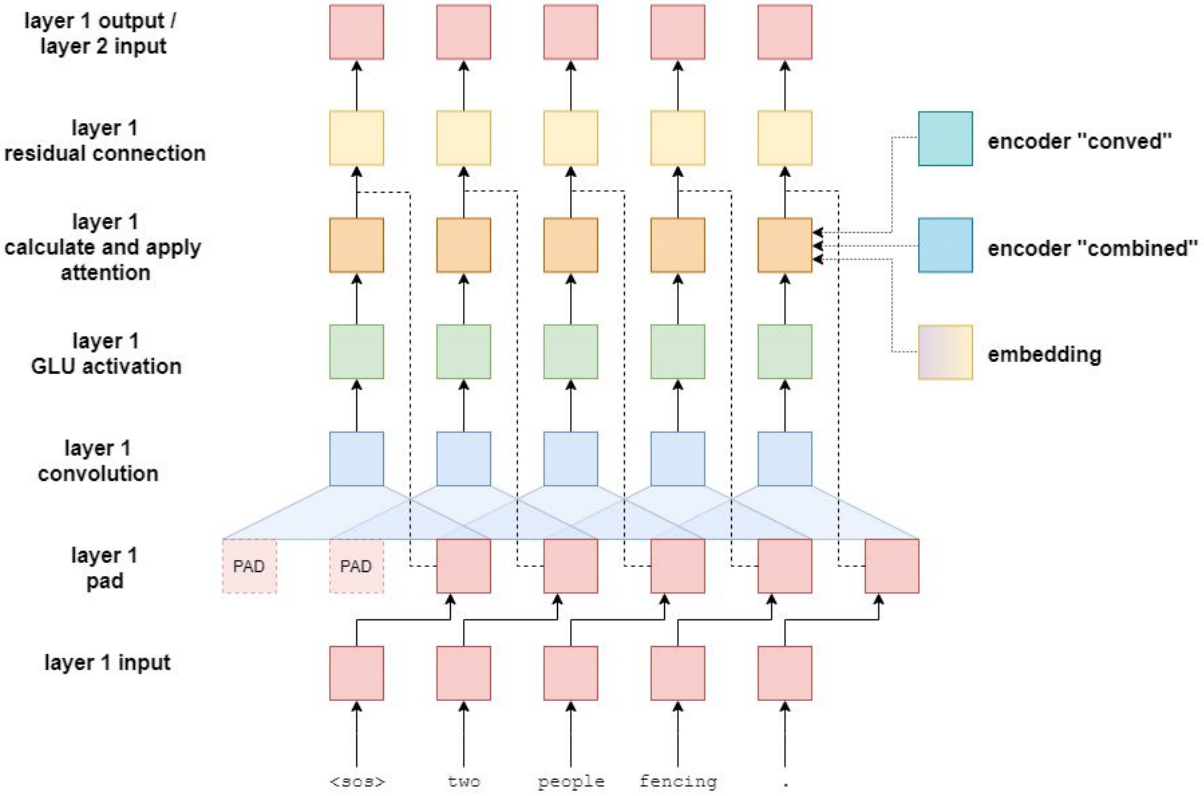
Conv seq2seq: Encoder



Conv seq2seq: Decoder



Conv seq2seq: Decoder

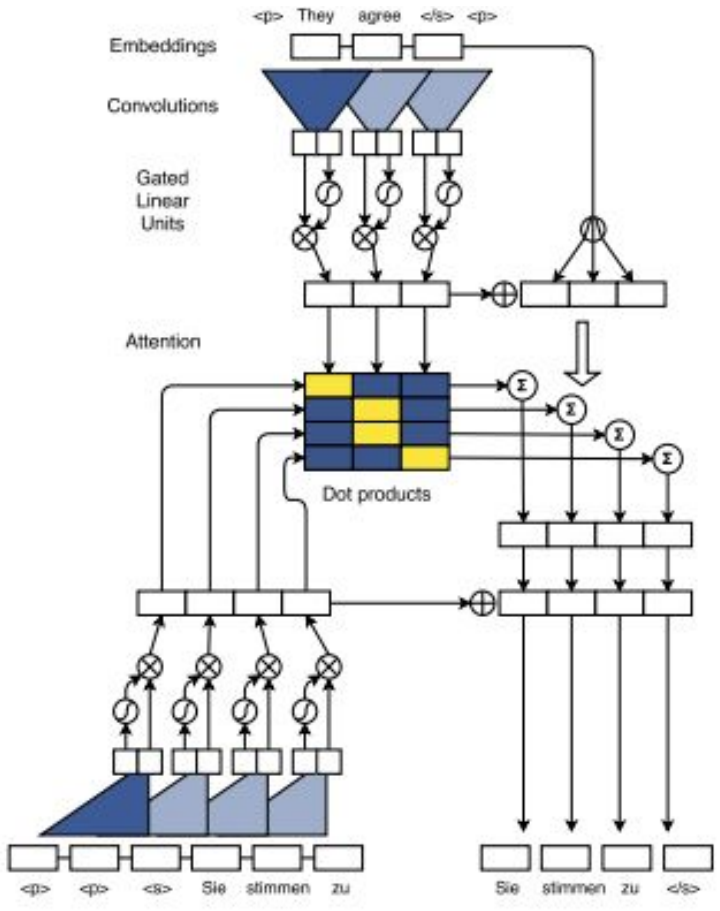


Conv seq2seq: summary

+ Легко параллелится
+ Качество (BLEU) оказалось выше
+ Контроль над параметрами
- Все равно обучается долго
- Сложнее интерпретировать

WMT'14 English-French	BLEU
Wu et al. (2016) GNMT (Word 80K)	37.90
Wu et al. (2016) GNMT (Word pieces)	38.95
Wu et al. (2016) GNMT (Word pieces) + RL	39.92
ConvS2S (BPE 40K)	40.51

WMT'14 English-German	BLEU
Luong et al. (2015) LSTM (Word 50K)	20.9
Kalchbrenner et al. (2016) ByteNet (Char)	23.75
Wu et al. (2016) GNMT (Word 80K)	23.12
Wu et al. (2016) GNMT (Word pieces)	24.61
ConvS2S (BPE 40K)	25.16



Summary

Итак, мы узнали

- О том, как оценивается машинный перевод
- Какие вообще есть подходы к этому
- О нейросетевых подходах (RNN, CNN) с некоторыми деталями
- Что такое attention и как он применяется в данной задаче

Список источников

- **Benson Kutuku** et al. Review on Machine Translation Approaches
- **Papineni** et al. BLEU
- **Nagao** Example-based MT
- Презентация **Abigail See** (Stanford University)
- RNN Cheatsheet
- **Sutskever** et al. Seq2seq Learning with Neural Networks
- **Cho** et al. Learning Phrase Representations using RNN Encoder-Decoder
- **Luong** et al. Attention-based NMT
- **Gehring** et al. Convolutional seq2seq
- Jupyter notebook - CNN seq2seq

Вопросы

1. Расскажите, как можно оценивать машинный перевод. Приведите пример с формулой и поясните идею стоящую за ней.
2. Расскажите про некоторые классические методы машинного перевода (как минимум 3). Какие у них есть преимущества и недостатки?
3. Как можно решать задачу машинного перевода с помощью RNN? Предложите архитектуру и идею, стоящую за ней. Назовите основные проблемы модели и способы борьбы с ними.
4. Расскажите про attention: что это такое, как применяется в RNN и какова мотивация для использования.