

Машинный перевод

Бакиева Аделина. БПМИ-171

Содержание

- Постановка задачи
 - Метрики
- До нейронных сетей
 - RBMT
 - Виды RBMT
 - EBMT
 - SMT
 - Виды SMT
- NMT
 - Seq2seq
 - Attention
 - Проблемы NMT

x – предложение на языке А

y – предложение на языке Б

$$MT(x) = y$$

Хотим при этом, чтобы x и y «хорошо» соотносились

Постановка задачи

Метрики

Человеческие

- Adequacy
(имеет ли предложение на выходе тот же смысл, что и исходное)
- Fluency
(насколько «похож» перевод на речь носителя языка)

Машинные

- Precision-Recall
(сравниваем предложение с эталоном по множеству слов)

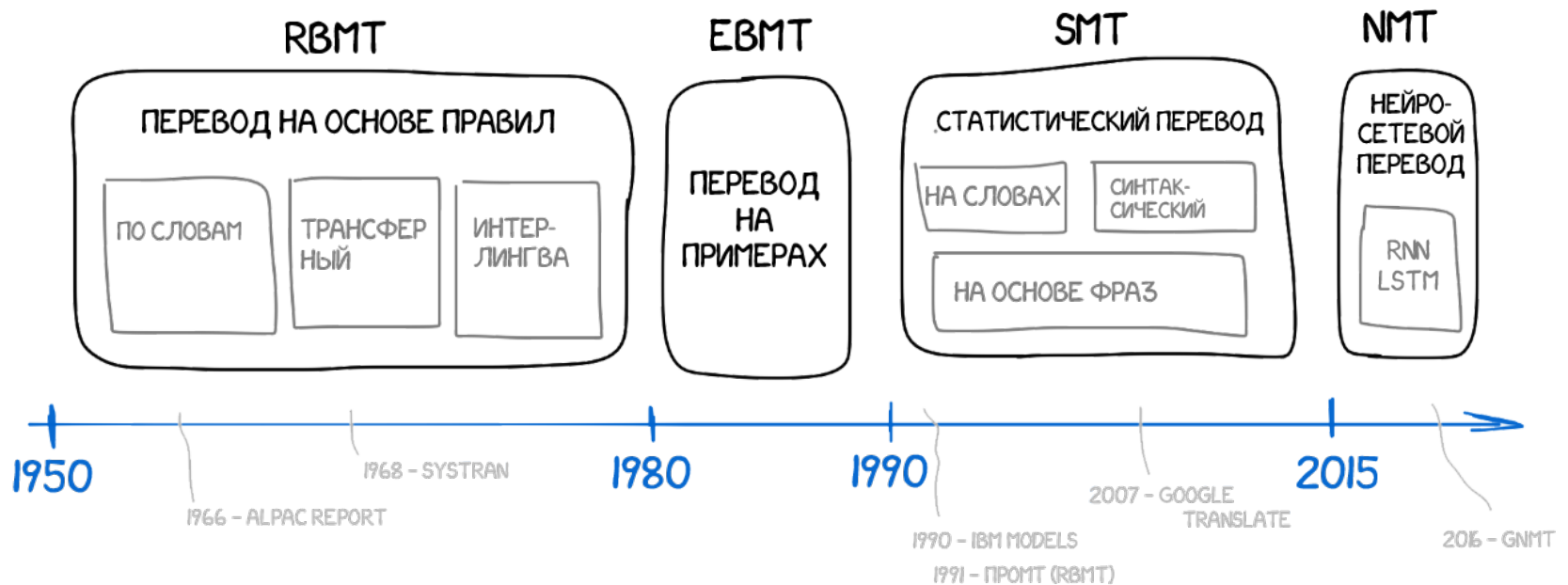
- Word Error Rate
$$\frac{\text{replacements} + \text{indertions} + \text{deletions}}{|\text{reference}|}$$

- BLEU

$$\min\left(1, \frac{|\text{output}|}{|\text{reference}|}\right) \left(\prod_{i=1}^4 \text{presicion}_i\right)^{\frac{1}{4}}$$

- ...

КРАТКАЯ ИСТОРИЯ МАШИННОГО ПЕРЕВОДА



До нейронных сетей

Переводим по словарю, затем с помощью правил



RBMT

Конвертируем в
интерлингву,
потом обратно

Интерлингвальный
перевод

Трансферный перевод

Выделяем синтаксические
конструкции, снова
используем правила для
согласования

Дословный перевод

Переводим
дословно, потом
используем правила
для согласования


Находим похожую фразу в базе → переводим →
объединяем компоненты → согласовываем части между
собой

(УЖЕ ЗНАКОМЫЙ НАМ ПРИМЕР)

Я ИДУ В ТЕАТР = I'M GOING TO THE THEATER

Я ИДУ В МАГАЗИН ^{???} = I'M GOING TO THE STORE

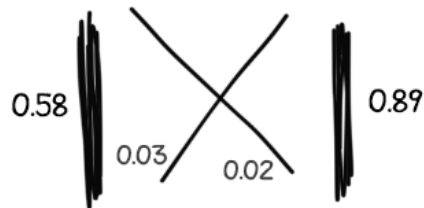
STORE



SMT

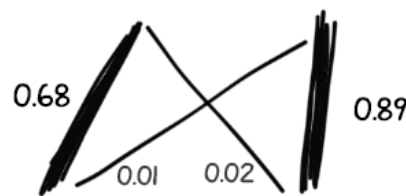
translation = $\text{argmax}_{\text{option}} P(\text{sentence} | \text{option}) P(\text{option})$

THE HOUSE



DAS HAUS

BLUE HOUSE



BLAUES HAUS

THE CAR



DAS AUTO

SMT

По словам

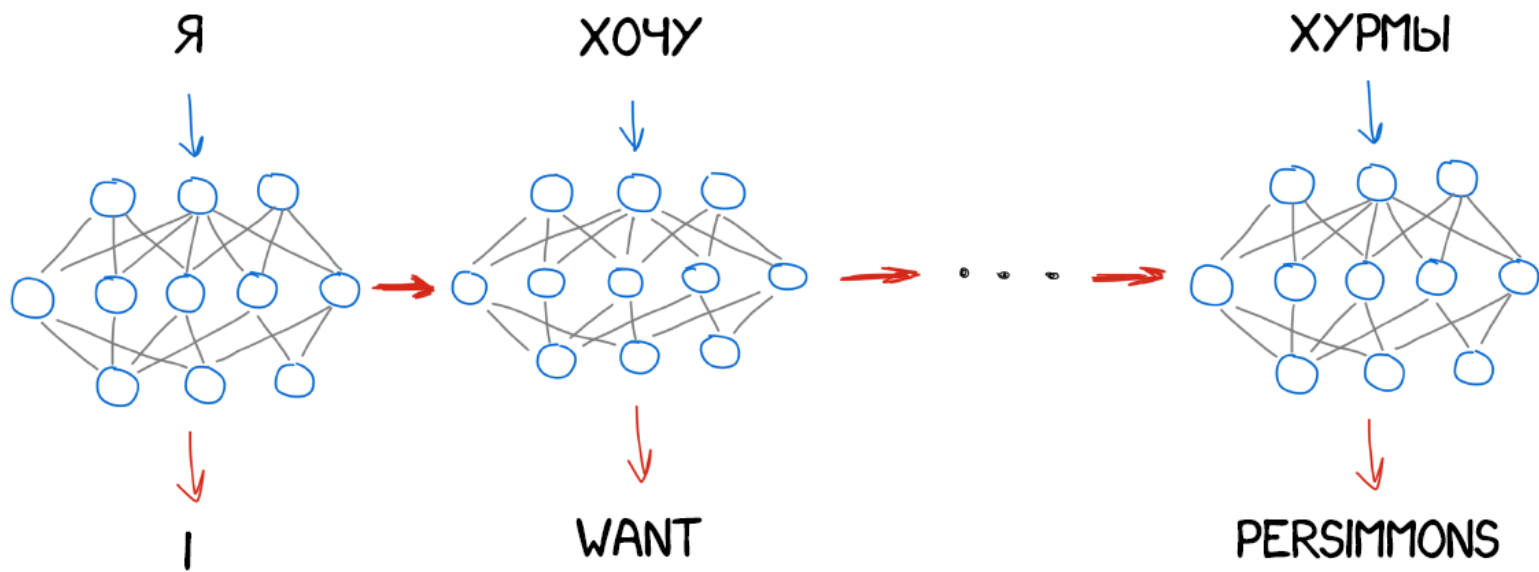
Пословный подбор
наиболее вероятного
варианта

Разбиваем текст
на N-граммы

По фразам

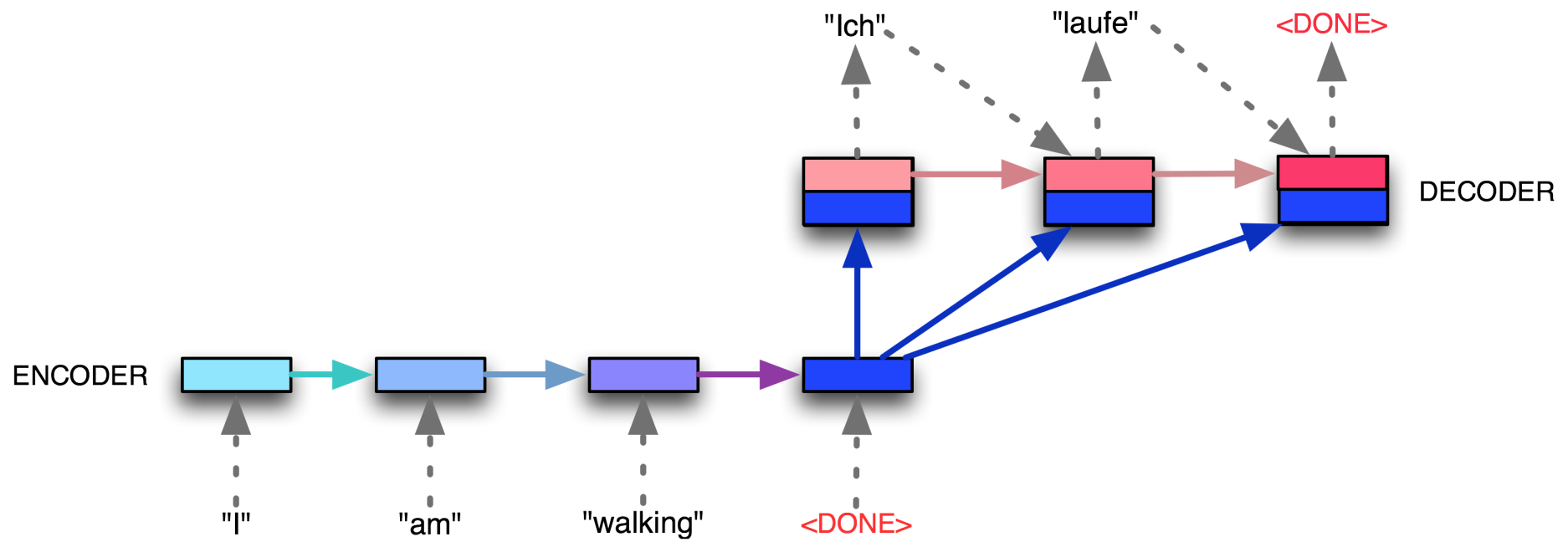
Синтаксический

Разбираем предложение
на синтаксические
конструкции, переводим
пословно, собираем
обратно по дереву



NMT

seq2seq

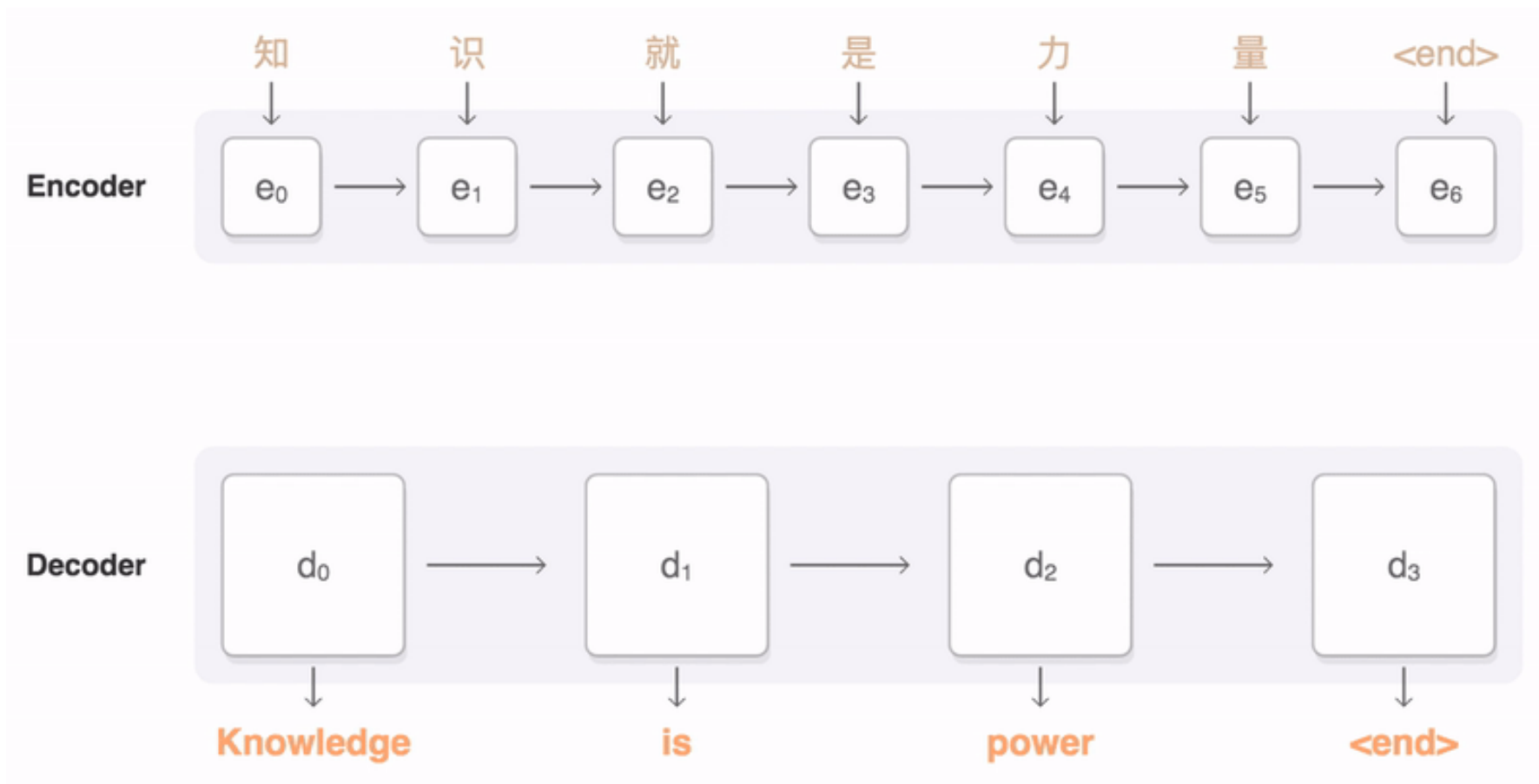


проблема

“В 1800-х годах, в те времена, когда не было еще ни железных, ни шоссейных дорог, ни газового, ни стеаринового света, ни пружинных низких диванов, ни мебели без лаку, ни разочарованных юношей со стеклышками, ни либеральных философов-женщин, ни милых дам-камелий, которых так много развелось в наше время, - в те наивные времена, когда из Москвы, выезжая в Петербург в повозке или карете, брали с собой целую кухню домашнего приготовления, ехали восемь суток по мягкой, пыльной или грязной дороге и верили в пожарские котлеты, в валдайские колокольчики и бублики, - когда в длинные осенние вечера нагорали сальные свечи, освещая семейные кружки из двадцати и тридцати человек, на балах в канделябры вставлялись восковые и спермацетовые свечи, когда мебель ставили симметрично, когда наши отцы были еще молоды не одним отсутствием морщин и седых волос, а стрелялись за женщин и из другого угла комнаты бросались поднимать нечаянно и не нечаянно уроненные платочки, наши матери носили коротенькие талии и огромные рукава и решали семейные дела выниманием билетиков, когда прелестные дамы-камелии прятались от дневного света, - в наивные времена масонских лож, мартинистов, тугендбунда, во времена Милорадовичей, Давыдовых, Пушкиных, - в губернском городе К. был съезд помещиков, и кончались дворянские выборы.”

Л.Н.Толстой

attention



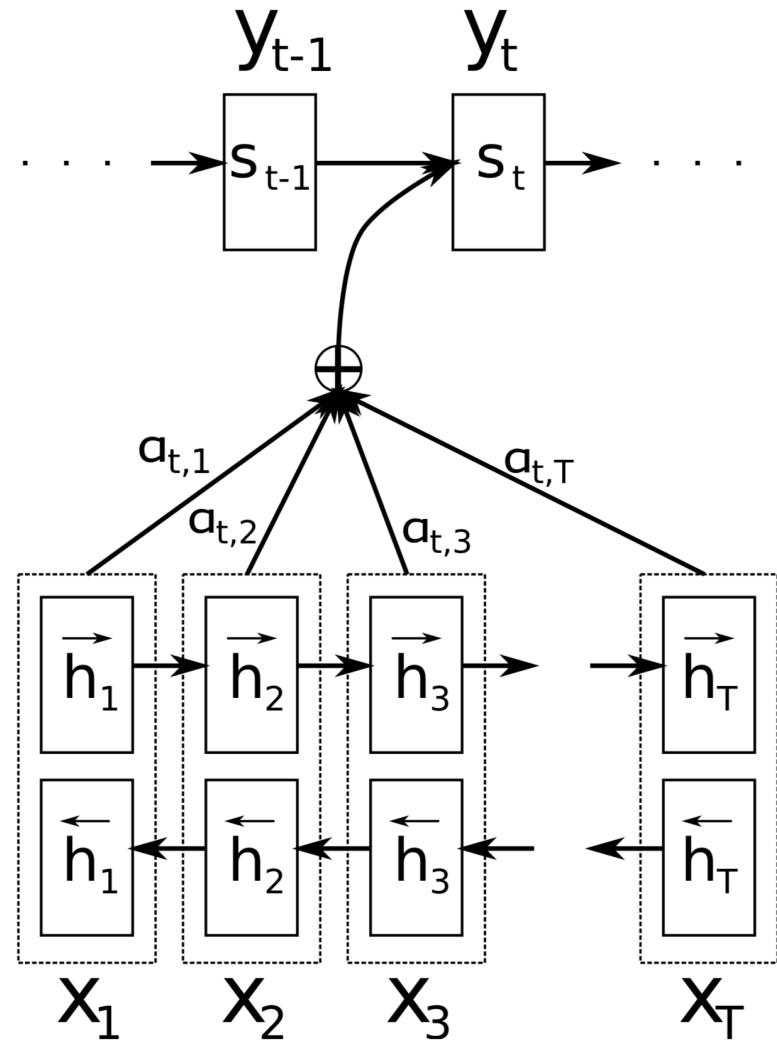
attention

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

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

$$\alpha_{ij} = \frac{\exp e_{ij}}{\sum_{k=1}^{T_X} \exp e_{ik}}$$

$$e_{ij} = a(s_{i-1}, h_j)$$



Проблемы NMT

Несоответствие корпусов

Если обучать на одном специфическом корпусе, а тестировать на другом, то качество сильно снижается

Длинные предложения

Несмотря на использование внимания, на предложениях длины ≥ 60 побеждает SMT. На длине ≥ 80 разница становится существенной

Alignment

Иногда они не соответствуют нашим представлениям о том, что должно было получиться

Размер данных для обучения

Если корпус для обучения не очень большой, то на нем выигрывает SMT

Beam Search

Начиная с какого-то момента, рассмотрение бóльшего количества возможных переводов уменьшает качество

Редкие слова

NMT проигрывает SMT на редких словах

Вопросы

- Метрика BLEU (BiLingial Evaluation Understudy)
- Какую проблему помогает обойти использование attention в seq2seq?
- Три проблемы NMT



Ссылки

https://vas3k.ru/blog/machine_translation/

<http://www.statmt.org/book/slides/08-evaluation.pdf>

https://hpi.de/fileadmin/user_upload/fachgebiete/plattner/teaching/MachineTranslation/MT2016/MT13_ExampleBasedMT.pdf

<https://habr.com/en/company/yandex/blog/224445/>

<https://guillaumegenthial.github.io/sequence-to-sequence.html>

<https://google.github.io/seq2seq/>

http://www.machinelearning.ru/wiki/images/c/cd/2017_417_PolykovskyDA.pdf

<https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf>

<https://arxiv.org/pdf/1409.0473.pdf>

<https://medium.com/@devnag/seq2seq-the-clown-car-of-deep-learning-f88e1204dac3>

<https://arxiv.org/pdf/1706.03872.pdf>

<https://medium.com/@ozinkegliyin/six-challenges-for-neural-machine-translation-8a780ead92ab>