

Машинный перевод

Machine Translation

Екатерина Владимировна Еникеева

12 октября 2022

Автоматическая обработка естественного языка, лекция 3

Зачем говорим об МТ?

- Одна из первых задач NLP
- Понятное приложение для конечного пользователя
- Многие методы NLP могут использоваться в МТ
- Методы МТ могут применяться в других задачах NLP

Проблемы МТ

➤ Неоднозначность

➤ на разных уровнях:

*The animal didn't cross the road because **it** was too wide.*

*The animal didn't cross the road because **it** was too tired.*

A computer that understands you like your mother

➤ + учёт контекста за рамками предложения

➤ + синонимия

➤ Разнообразие естественных языков

➤ порядок слов, морфология ...

Приложения МТ

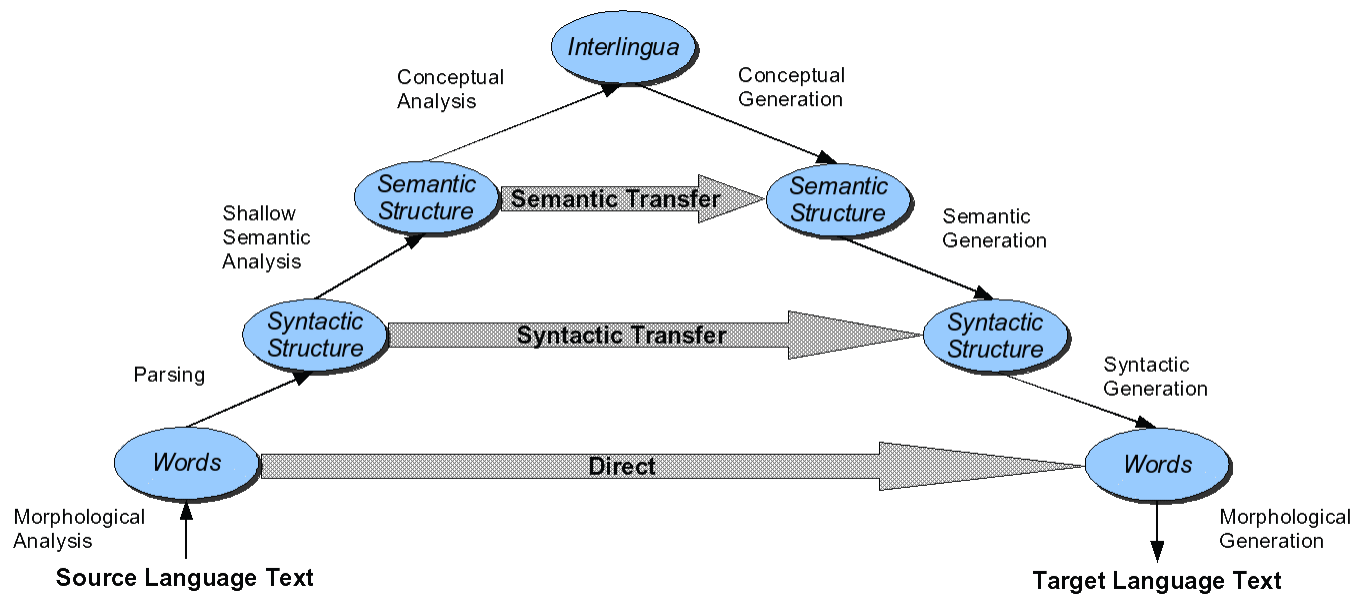
- Автоматический перевод
 - Online переводчики
 - Domain-specific перевод (патенты, локализация)
 - Перевод в мессенджерах
 - ...
- Инструменты переводчиков
 - computer-assisted translation, CAT
 - translation memory – переводческая память
- Оценка качества перевода / MT quality evaluation

Ранние эксперименты

- 1946-48 Warren Weaver, идея автоматического перевода с использованием словаря фраз
- 1949 Weaver memorandum
- 1954 IBM/Georgetown – Джорджтаунский эксперимент
- 1958 первая Всесоюзная конференция по МП
- 1955-65 развитие МП сразу в нескольких лабораториях в США
- 1966 доклад ALPAC, эксперименты сворачиваются

<http://www.hutchinsweb.me.uk/PPF-TOC.htm>

Перевод через интерлингву



Rule-based модели

- Direct translation – прямой перевод:
 - Словари
 - Маппинг грамматики
- Transfer model
 - Парсим входной текст
 - Применяем правила (transfer) и преобразуем в дерево языка перевода
 - Генерируем предложение по дереву

Терминология

Source language – **f** (foreign) / src

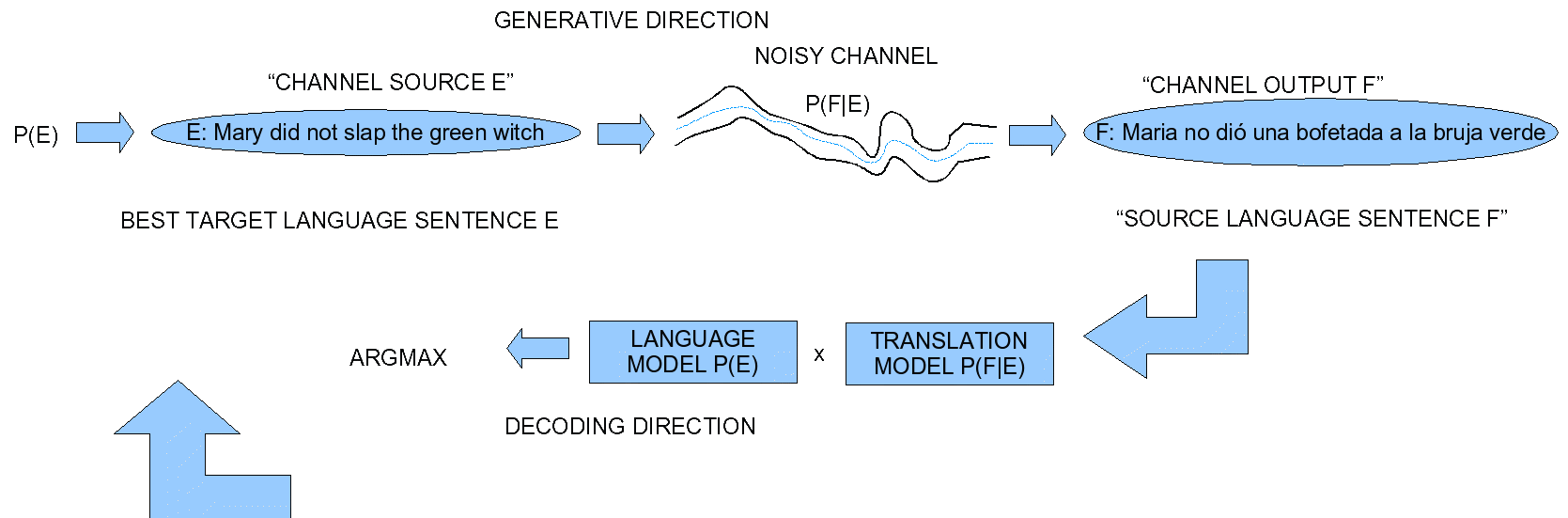
Destination / target language – **e** (English / input) / dst

Два основных параметра:

- fluency
- adequacy / faithfulness

Ручная оценка качества обычно смещена в сторону
fluency

Noisy channel model



Статистический МП

$$\begin{aligned}\hat{E} &= \operatorname{argmax}_{E \in \text{English}} P(E | F) &= \operatorname{argmax}_{E \in \text{English}} \frac{P(F | E)P(E)}{P(F)} \\ & &= \operatorname{argmax}_{E \in \text{English}} P(F | E)P(E)\end{aligned}$$

Translation ModelLanguage Model

Translation Model – модель перевода / фразовая таблица
Language Model – языковая модель

Peter Brown, Stephen A. Della Pietra,
Vincent J. Della Pietra, Robert L. Mercer.
1993. The Mathematics of Statistical
Machine Translation: Parameter
Estimation. Computational Linguistics
19:2, 263-311. **“The IBM Models”**

Статистическая языковая модель

Модель порядка K

$$P(w_1 w_2 \dots w_n) \approx \prod_i P(w_i \mid w_{i-k} \dots w_{i-1})$$

$$P(w_i \mid w_1 w_2 \dots w_{i-1}) \approx P(w_i \mid w_{i-k} \dots w_{i-1})$$

Оценка вероятности по частотам
(count-based):

$$P(w_i) = \frac{\text{count}(w_i)}{\sum_{w \in C} \text{count}(w)}$$

Статистическая модель перевода

bag	сумка	0.3
bag	мешок	0.13
bag of words	мешок слов	0.05
bag of	пакет с	0.2

Откуда берем эти вероятности?

- (скорее всего) count-based подход
- нужно знать частоты слов / фраз из **e** и **f**
- нужно знать, когда эти фразы являются переводами друг друга

➤ Модель перевода – фразовая таблица (phrase table / PT)

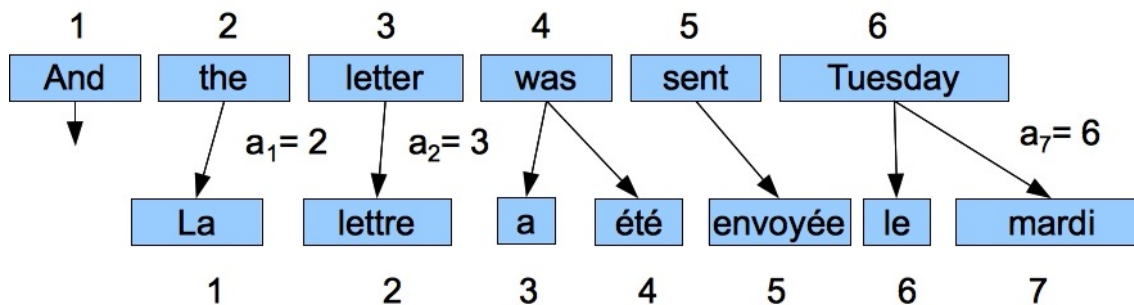
Обучающие данные

- Для языковой модели – большой корпус языка перевода – *monolingual corpus*
- Для модели перевода – большой параллельный корпус (bibtex):
 - Выравнивание по предложениям
 - Выравнивание по словам

Параллельные корпуса

- [EuroParl](#)
- [ParaCrawl](#)
- [OPUS](#) : коллекция корпусов разных доменов
- CC Matrix
- ...

Выравнивание / alignment



$A = [2, 3, 4, 4, 5, 6, 6]$

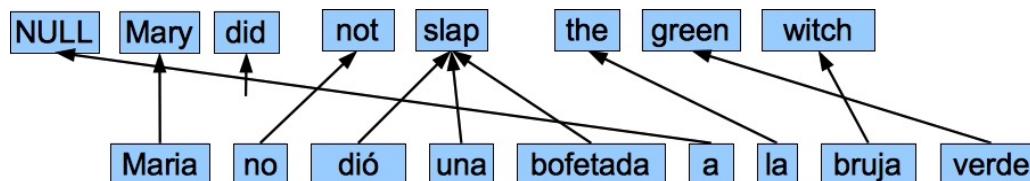
*NB! Записываем из **f** в **e***

	Le	lettre	a	été	envoyée	le	Mardi
And							
the							
letter							
was							
sent							
Tuesday							

Выравнивание

Предположения:

- Выравнивание по предложениям уже есть
- Строим отображение один – много
(то есть одно слово из **f** может отображаться
несколько из **e**, но не наоборот)
- Есть нулевой элемент
(слово из **f** может не иметь переводного эквивалента
в **e**)



IBM Model 1

IBM Model 1 (есть ещё 2, ...) —

простая генеративная модель, описывает, как мы получаем F , имея $E=e_1, e_2, \dots, e_I$

- Пусть J — количество слов в F : $F=f_1, f_2, \dots, f_J$
- Построим выравнивание $A=a_1, a_2, \dots, a_J$
- Для каждой позиции j в F , генерируем слово f_j из слова в E : e_{a_j}

IBM Model 1a

- Пусть $t(f_x, e_y)$ — вероятность перевода e_y в f_x
- Если мы знаем предложение E , выравнивание A и длину входа J , то можем посчитать вероятность исходного предложения

$$P(F | E, A) = \prod_{j=1}^J t(f_j, e_{a_j})$$

- Зачем нам это? Нам нужно такое выравнивание, которое максимизирует эту вероятность

IBM Model 1

Вероятность такого события:

Предложение **f** переводится в **e** и выравнивается функцией *a*

$$p(\mathbf{e}, a|\mathbf{f}) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

l – длины,

$t(e|f)$ – вероятность пословного перевода,

$(l_f + 1)^{l_e}$ – число возможных выравниваний,

ϵ – нормировка

IBM Model 1 probabilities

Как оценить вероятность выравнивания?

$$p(a|\mathbf{e}, \mathbf{f}) = \frac{p(\mathbf{e}, a|\mathbf{f})}{p(\mathbf{e}|\mathbf{f})}$$

Числитель – умеем выводить из пословных вероятностей t

В любом случае, надо оценить знаменатель – вероятность перевода \mathbf{f} в \mathbf{e} при любом выравнивании

IBM Model 1 target

$$\begin{aligned} p(\mathbf{e}|\mathbf{f}) &= \sum_a p(\mathbf{e}, a|\mathbf{f}) \\ &= \sum_{a(1)=0}^{l_f} \dots \sum_{a(l_e)=0}^{l_f} p(\mathbf{e}, a|\mathbf{f}) = \sum_{a(1)=0}^{l_f} \dots \sum_{a(l_e)=0}^{l_f} \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j | f_{a(j)}) \\ &= \frac{\epsilon}{(l_f + 1)^{l_e}} \sum_{a(1)=0}^{l_f} \dots \sum_{a(l_e)=0}^{l_f} \prod_{j=1}^{l_e} t(e_j | f_{a(j)}) \\ &= \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j | f_i) \end{aligned}$$

IBM Model 1 : E-step

$$\begin{aligned} p(a|\mathbf{e}, \mathbf{f}) &= \frac{p(\mathbf{e}, a|\mathbf{f})}{p(\mathbf{e}|\mathbf{f})} \\ &= \frac{\frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})}{\frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} \sum_{i=0}^{l_f} t(e_j|f_i)} \\ &= \prod_{j=1}^{l_e} \frac{t(e_j|f_{a(j)})}{\sum_{i=0}^{l_f} t(e_j|f_i)} \end{aligned}$$

Сводим
предыдущие
формулы
вместе, чтобы
таким образом
получить
вероятность
выравнивания

IBM Model 1 : M-step

Что мы можем оценить по данным?

δ – функция Кронекера: $\delta(x, y) = 1$, если $x=y$, иначе 0

$$c(e|f; \mathbf{e}, \mathbf{f}) = \frac{t(e|f)}{\sum_{i=0}^{l_f} t(e|f_i)} \sum_{j=1}^{l_e} \delta(e, e_j) \sum_{i=0}^{l_f} \delta(f, f_i)$$

$$t(e|f; \mathbf{e}, \mathbf{f}) = \frac{\sum_{(\mathbf{e}, \mathbf{f})} c(e|f; \mathbf{e}, \mathbf{f})}{\sum_e \sum_{(\mathbf{e}, \mathbf{f})} c(e|f; \mathbf{e}, \mathbf{f})}$$


IBM Model 1 : Псевдокод

```
Input: set of sentence pairs (e, f)
Output: translation prob.  $t(e|f)$ 
1: initialize  $t(e|f)$  uniformly
2: while not converged do
3:   // initialize
4:   count( $e|f$ ) = 0 for all  $e, f$ 
5:   total( $f$ ) = 0 for all  $f$ 
6:   for all sentence pairs (e, f) do
7:     // compute normalization
8:     for all words  $e$  in e do
9:       s-total( $e$ ) = 0
10:    for all words  $f$  in f do
11:      s-total( $e$ ) +=  $t(e|f)$ 
12:    end for
13:  end for
14:  // collect counts
15:  for all words  $e$  in e do
16:    for all words  $f$  in f do
17:      count( $e|f$ ) +=  $\frac{t(e|f)}{s\text{-total}(e)}$ 
18:      total( $f$ ) +=  $\frac{t(e|f)}{s\text{-total}(e)}$ 
19:    end for
20:  end for
21: end for
22: // estimate probabilities
23: for all foreign words  $f$  do
24:   for all English words  $e$  do
25:      $t(e|f) = \frac{\text{count}(e|f)}{\text{total}(f)}$ 
26:   end for
27: end for
28: end while
```



IBM Model 1 : Пример

IBM 1 – модель,
использующая
только t – **lexical**
translation
probability


das Haus
the house



das Buch
the book



ein Buch
a book



e	f	Initial	1st it.	2nd it.	3rd it.	...	Final
<i>the</i>	<i>das</i>	0.25	0.5	0.6364	0.7479	...	1
<i>book</i>	<i>das</i>	0.25	0.25	0.1818	0.1208	...	0
<i>house</i>	<i>das</i>	0.25	0.25	0.1818	0.1313	...	0
<i>the</i>	<i>buch</i>	0.25	0.25	0.1818	0.1208	...	0
<i>book</i>	<i>buch</i>	0.25	0.5	0.6364	0.7479	...	1
<i>a</i>	<i>buch</i>	0.25	0.25	0.1818	0.1313	...	0
<i>book</i>	<i>ein</i>	0.25	0.5	0.4286	0.3466	...	0
<i>a</i>	<i>ein</i>	0.25	0.5	0.5714	0.6534	...	1
<i>the</i>	<i>haus</i>	0.25	0.5	0.4286	0.3466	...	0
<i>house</i>	<i>haus</i>	0.25	0.5	0.5714	0.6534	...	1

IBM Models (4)

Про оценку параметров:

<http://mt-class.org/jhu/slides/lecture-ibm-model1.pdf>

(слайды Р.Коэна)

Книжка по SMT:

Philipp Koehn. 2009. Statistical Machine Translation. Cambridge University Press

<https://www.cambridge.org/core/books/statistical-machine-translation/94EADF9F680558E13BE759997553CDE5>

IBM Models (5)

Что дальше?

- Перестановки / reordering
- Вставки-удаления / fertility
- Классы слов / word classes
- ...

Декодинг (1)

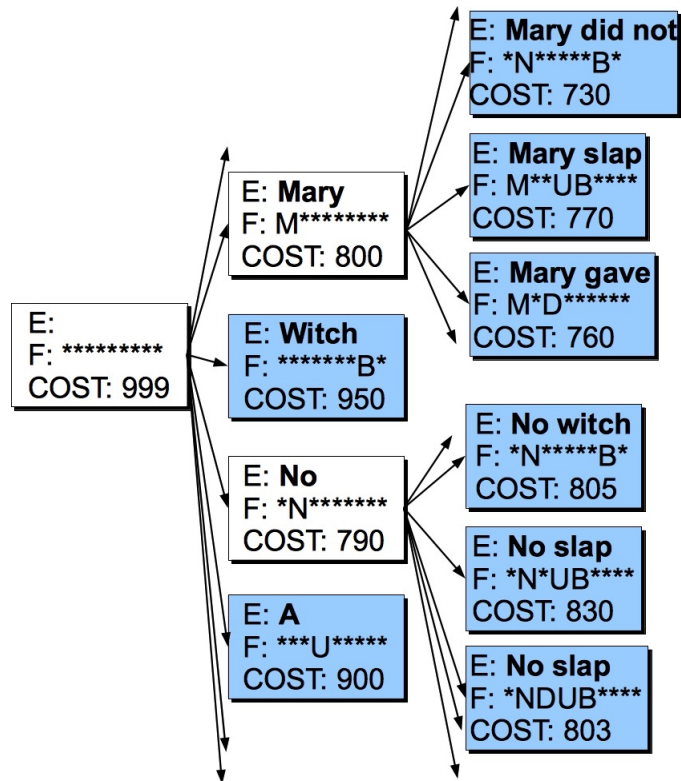
Как же получить перевод, если у нас есть РТ и LM?

Maria	no	dió	una	bofetada	a	la	bruja	verde
-------	----	-----	-----	----------	---	----	-------	-------

Mary	not	give	a	slap	to	the	witch	green
	did not		a slap	to			green witch	
	no		slap		to the			
	did not give				to			
					the			
			slap			the witch		

Stack-based decoding

Декодирование (2)



Decoding by hypothesis expansion

Слишком большое пространство поиска:

- Обрезаем гипотезы (pruning)
- Beam search (лучевой поиск?)

Пример графа поиска для большого предложения

Автоматическая оценка МТ

- Сравнение с эталоном — много метрик
- **BLEU** – учитывает точность по n-граммам и штрафует большую разницу в длине

$$\text{precision}_n = \frac{\sum_{C \in \text{corpus}} \sum_{n\text{-gram} \in C} \text{count-in-reference}_{\text{clip}}(n\text{-gram})}{\sum_{C \in \text{corpus}} \sum_{n\text{-gram} \in C} \text{count}(n\text{-gram})}$$

$$\text{BLEU-4} = \min\left(1, \frac{\text{output-length}}{\text{reference-length}}\right) \prod_{i=1}^4 \text{precision}_i$$

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. Proceedings of ACL 2002

Автоматическая оценка МТ

- Многозначность / синонимия / ...
- Гладкость ?
- Зависимость от эталонов
 - Human Evaluation
 - Quality Estimation
- MQM – Multi-Dimensional Quality Metrics:
 - terminology
 - accuracy
 - linguistic conventions
 - ...

Neural Machine Translation

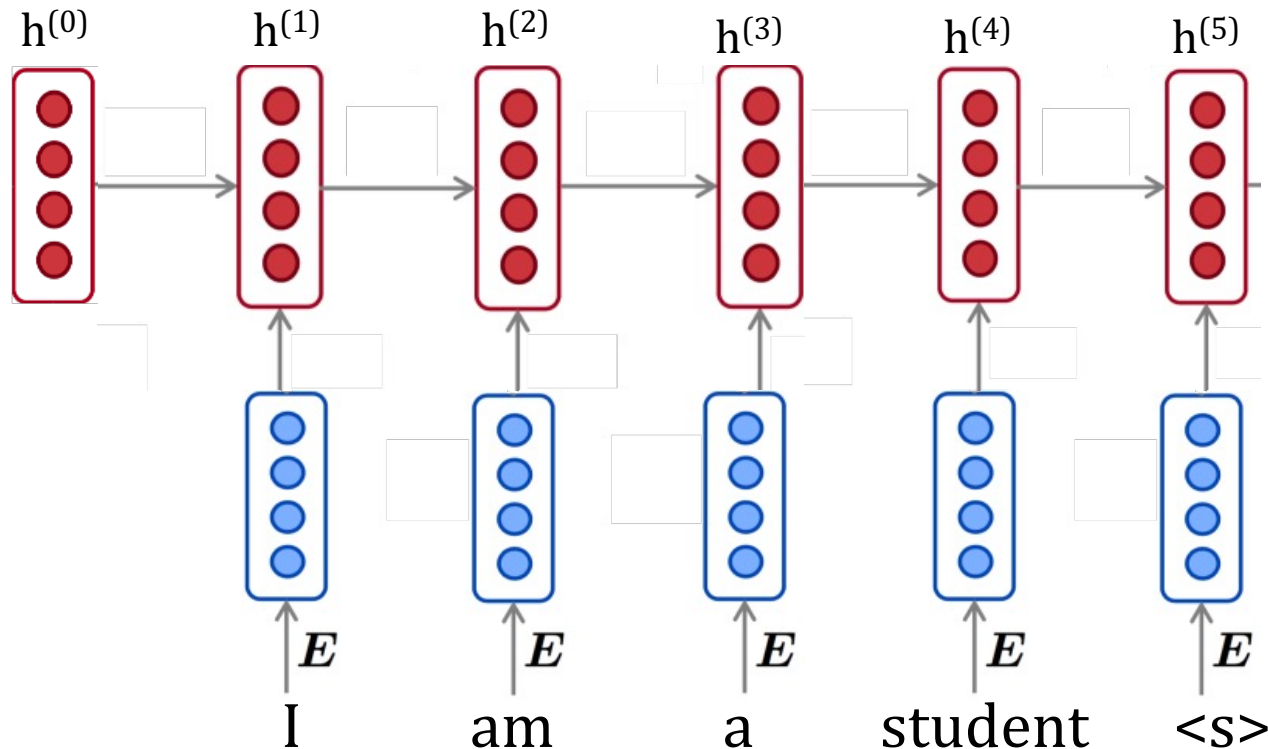
- MT – sequence-to-sequence task
- neural LMs
- encoding-decoding architecture:
 - encode source sentence -> embedding
 - decode into target sentence -> вероятности всех слов в словаре для каждой позиции
- Input: source + target pair,
output: вероятности всех слов в словаре для каждой позиции target'a

Preprocessing

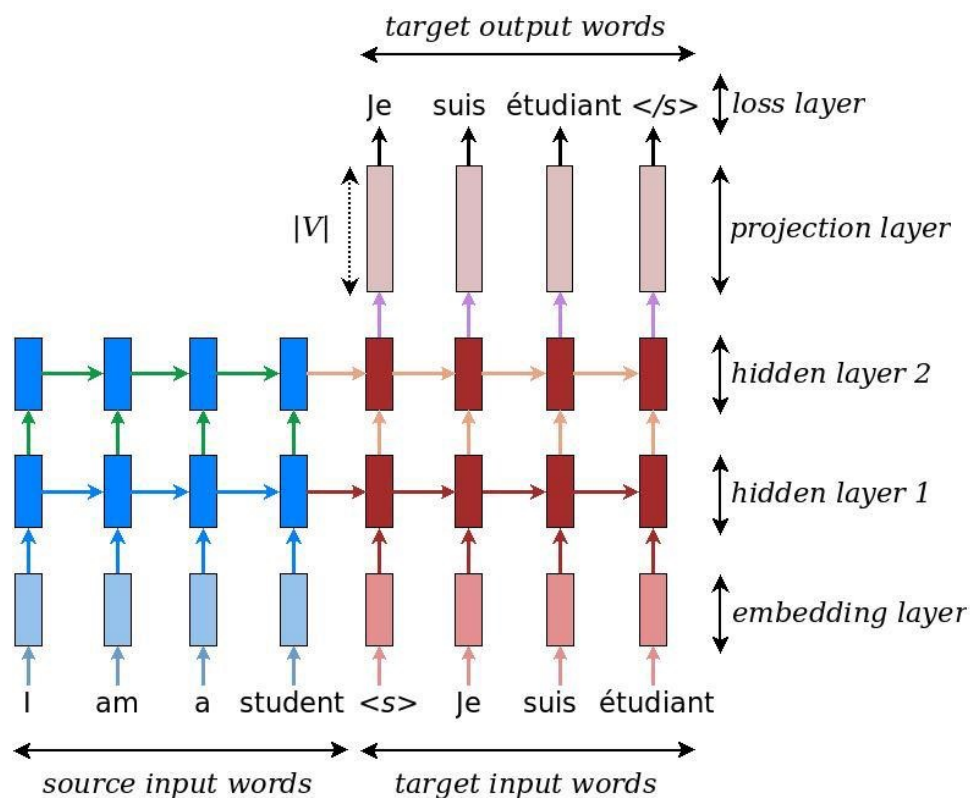
- Токенизация — subword units
BPE / wordpiece
- Sentence alignment
multilingual sentence embeddings
- Backtranslation

RNN reminder

- Language modeling



RNN-based architecture

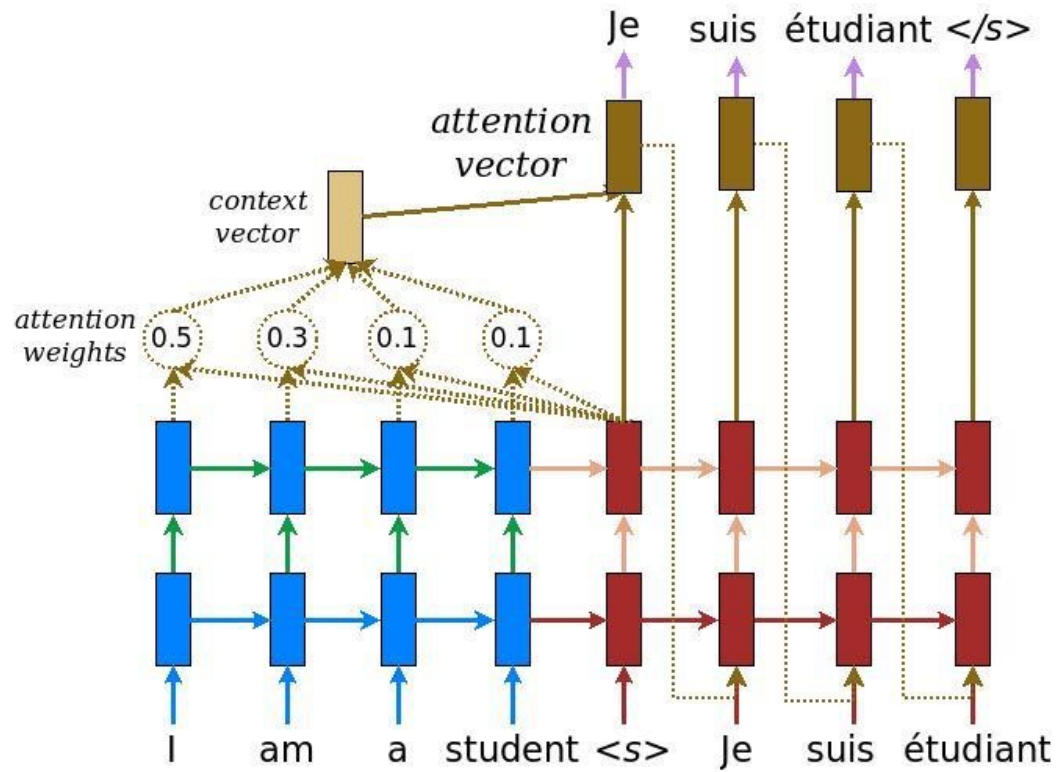


В чем проблема такой архитектуры для перевода?

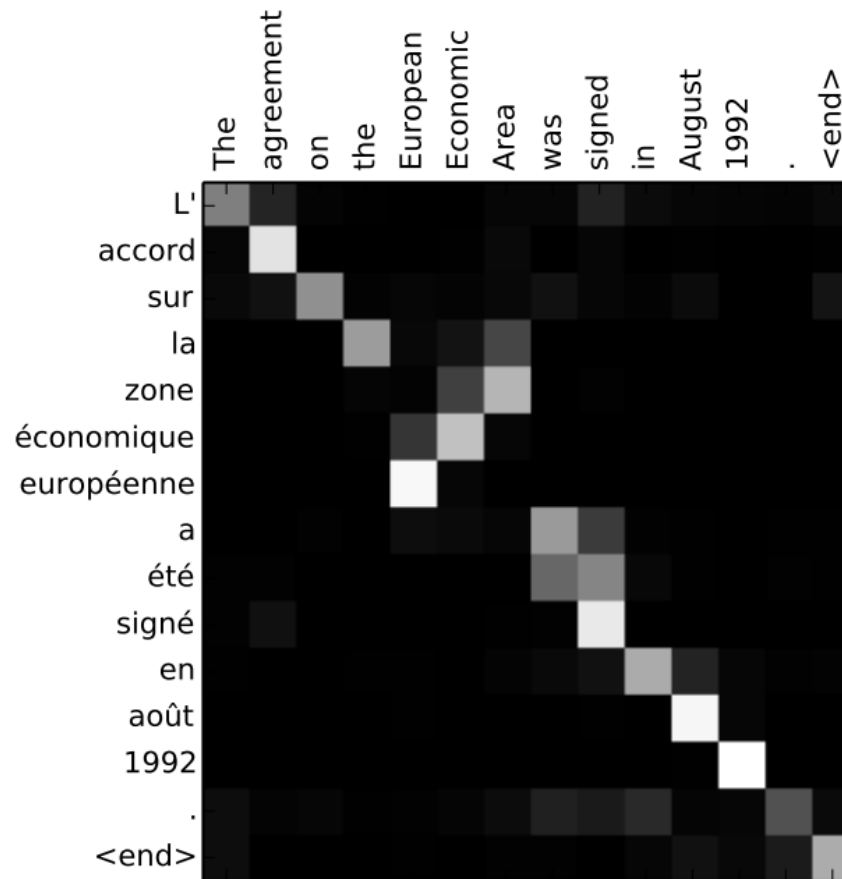
Attention is all you need

- Идея: давайте хранить скрытые представления для всех входных слов и пытаться учитывать только некоторые из них на каждом шаге декодера
- Скрытые представления: RNN в двух направлениях
- Теперь входное предложение – матрица \mathbf{H}
- Вводим новый параметр – вектор α , который при умножении на матрицу \mathbf{H} будет давать «вектор контекста» \mathbf{c}
- Вектор α будет показывать, насколько нужно учитывать каждое входное слово; получаем его с помощью состояния декодера \mathbf{h} (например, скалярное произведение)

Attention



Интерпретация внимания в МТ



Self-Attention

Вместо рекуррентных связей можем использовать контекст любого размера

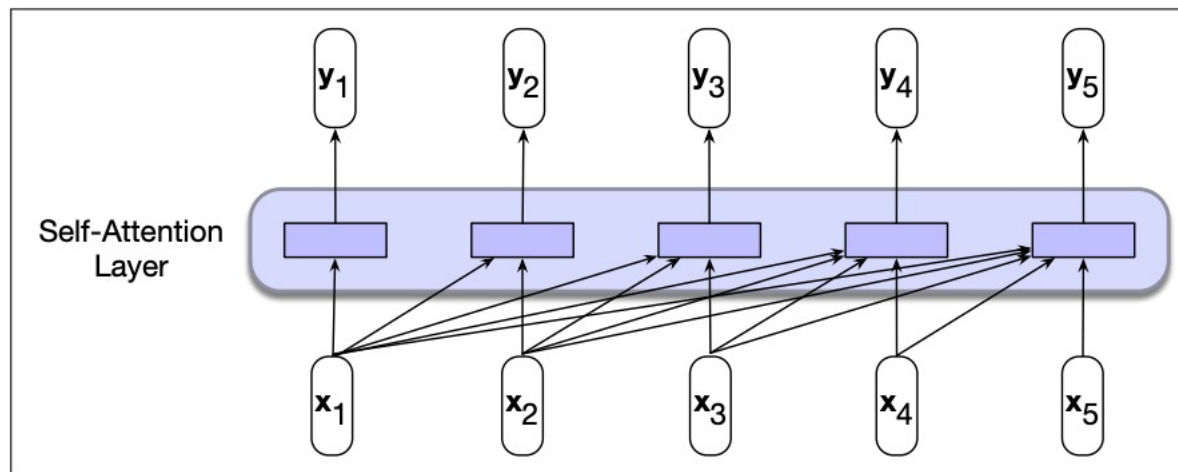


Figure 9.15 Information flow in a causal (or masked) self-attention model. In processing each element of the sequence, the model attends to all the inputs up to, and including, the current one. Unlike RNNs, the computations at each time step are independent of all the other steps and therefore can be performed in parallel.

Self-Attention

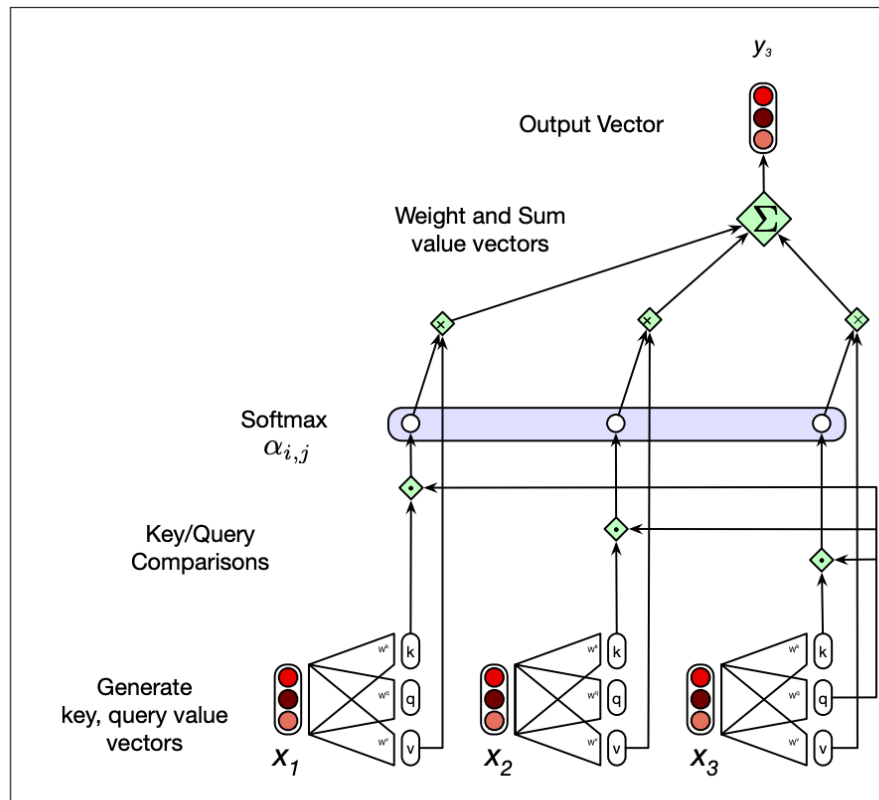


Figure 9.16 Calculating the value of y_3 , the third element of a sequence using causal (left-to-right) self-attention.

Трансформер

- **Self-Attention**
- **Multi-Head Attention** – несколько слоев self-attention, каждый со своими матрицами весов; затем объединяем их в один вектор с помощью ещё одной матрицы весов
- **Positional Encoding** — учитываем порядок

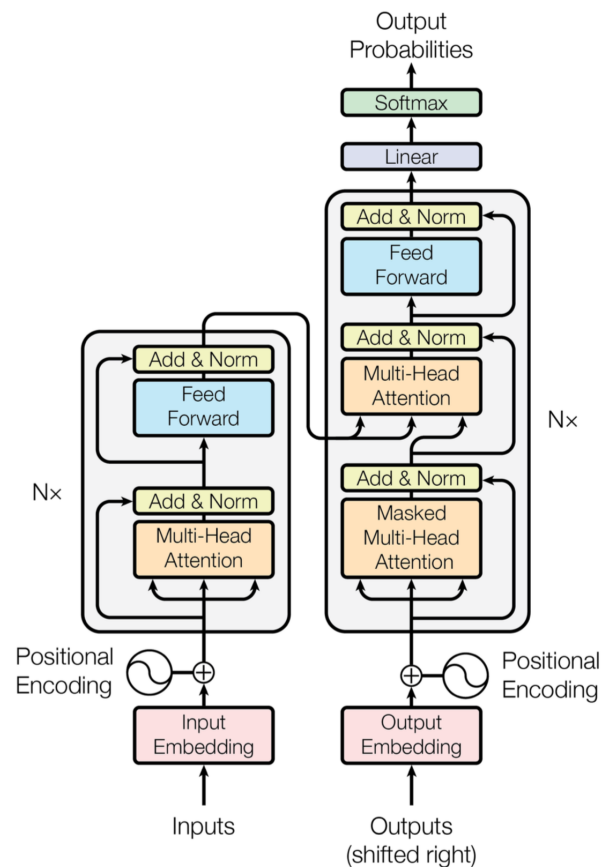


Figure 1: The Transformer - model architecture.

Issues

- Bias, ethical issues
- Low-resourced languages
- Domain-specific translation (glossaries)
- ...

Transfer Learning

- Pre-training
 - pretrained language models: обучаемся простой задаче, не требующей разметки, на больших данных
- Fine-tuning
 - используем параметры предобученной модели, чтобы дообучиться на других данных (например, более сложная разметка)
- Few-shot learning
 - GPT-3 etc.

Challenges

English→Russian

Ave.	Ave. z	System
91.8	0.681	HUMAN
81.5	0.469	Online-G
83.7	0.461	OPPO
79.6	0.404	ariel xv
80.3	0.336	Online-B
75.1	0.252	PROMT-NMT
76.2	0.222	DiDi-NLP
75.3	0.081	Online-A
71.3	0.035	zlabs-nlp
68.5	0.012	Online-Z

German→French

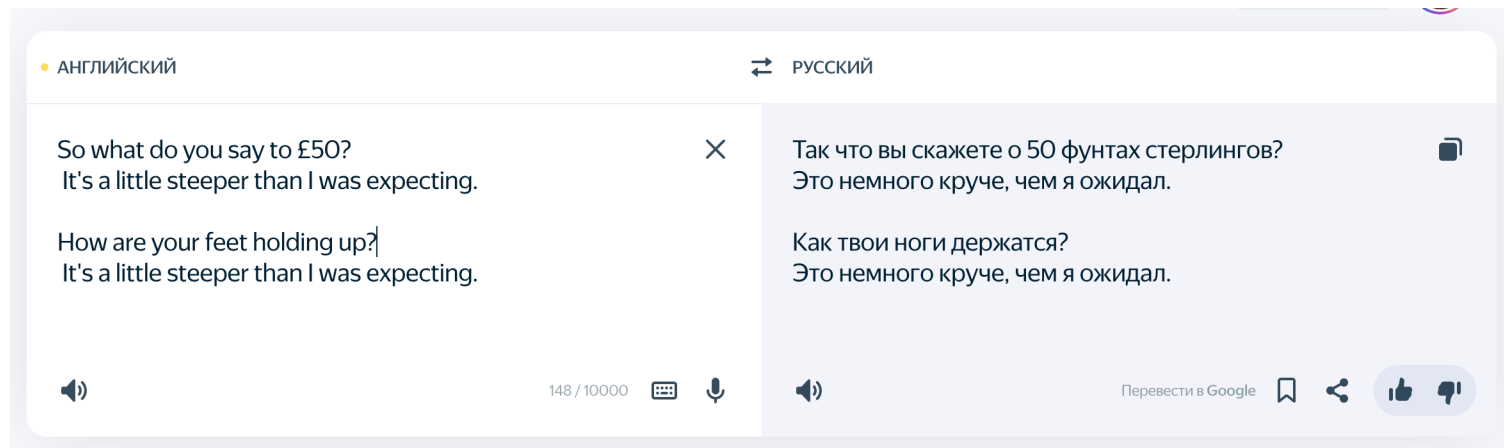
Ave.	Ave. z	System
90.4	0.279	OPPO
90.2	0.266	VolcTrans
89.7	0.262	IIE
89.2	0.243	HUMAN
89.1	0.226	Online-B
89.1	0.223	Online-A
88.5	0.208	Online-G

WMT 2020:

<http://statmt.org/wmt20/pdf/2020.wmt-1.1.pdf>

Challenges

- Multimodal MT:
 - изображения, звук, видео
- Контекстный перевод (discourse-aware translation), пример из статьи:



Применение SMT

Примеры использования

все

содержание

содержимое

контент

материал



The nickel **content** as well.

Равно как и **содержание** никеля.



It reduces it into a data stream which I am in the process of reverse-engineering to recreate the **content**.

Он сжимает его в поток данных, который я скоро разберу на кусочки, чтобы восстановить **содержимое**.



I know you'd all rather be at home, binge watching media **content**.

Я знаю, что вы все предпочли бы быть дома и без перерыва смотреть медиа- **контент**.



The paper will continue, but it'll offer **content** from a wider area, using pooled resources.

Газета останется, но будет освещать **материал** со многих мест, используя объединенные ресурсы.



What Corky means is that we can prove a rock is a meteorite simply by measuring its chemical **content**."

Корки говорит, мы можем доказать, что камень является метеоритом, просто исследовав его химический **состав**.



Instead of poverty, general prosperity and **content**; instead of hostility, harmony and unity of interests.

Словарь

content сущ

- содержание** ср **содержимое** ср **контент** м
наполнение ср **информационное наполнение**
substance, contained, filling, information content

- материал** м
material

- состав** м
composition

- довольство** ср
contentment

content гл

- довольствоваться** **удовольствоваться**
settle, be content

content прил

- довольный**
satisfied
- содержательный**
meaningful
- содержимый**
contained

<https://translate.yandex.ru/?lang=en-ru&text=content>

Что ещё почитать?

- WMT – ежегодный воркшоп и соревнование систем МП и оценки качества <http://statmt.org/wmt21/>
- Лучший в мире курс по NLP (ШАД)
https://github.com/yandexdataschool/nlp_course/
- Стэнфордский курс по NLP (есть ссылки на предыдущие годы)

<http://web.stanford.edu/class/cs224n/>

- сайт с библиографией по МП (после 2015 года не очень часто обновляется) <http://www.mt-archive.info/>

Спасибо!

Вопросы?