Машинный перевод

Высшая школа экономики
Научно-исследовательский семинар
“Машинное обучение и приложения”

Котов Егор, 172

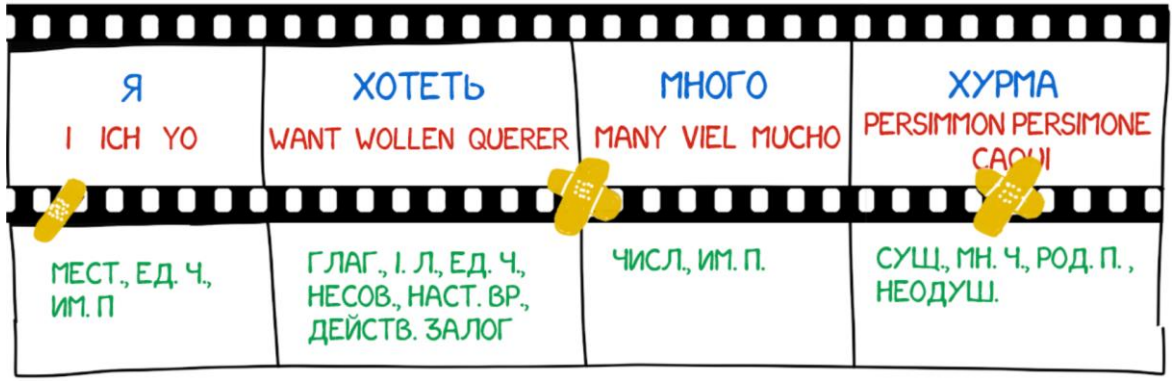
6 декабря 2019

План

1. Зачем нужен машинный перевод, история его развития
2. Простые методы машинного перевода
   - Перевод на основе правил
   - Перевод, основанный на примерах
   - Статистические модели
   - Сравнение методов
3. Нейросетевые подходы
   - Используемые метрики
   - Метод seq2seq с вниманием
   - Сравнение нейронного подхода со статистическим

История развития

Машина Петра Троянского, 1933

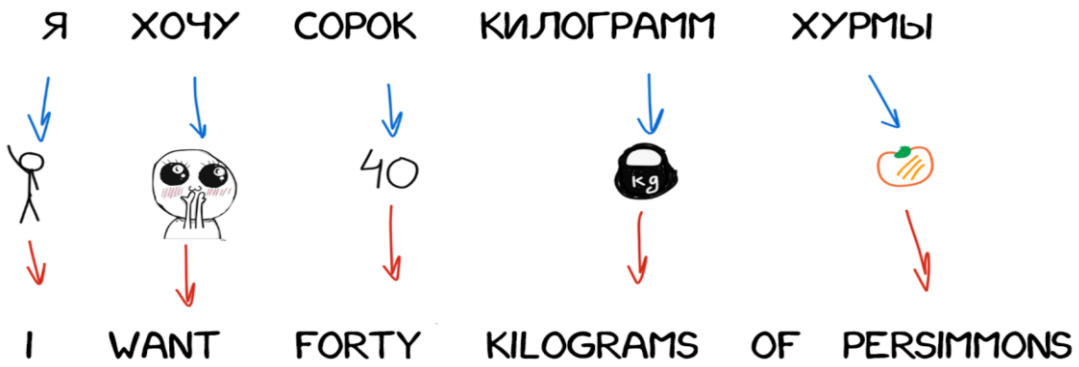


Джорджтаунский эксперимент, 1954

Перевод на основе правил

- Словарный подход

- Межъязыковой машинный перевод



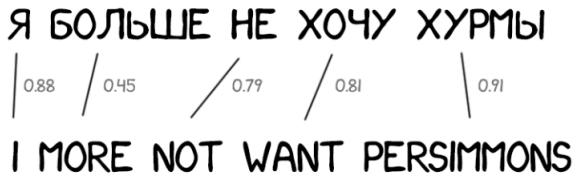
- Трансферные системы

Перевод, основанный на примерах

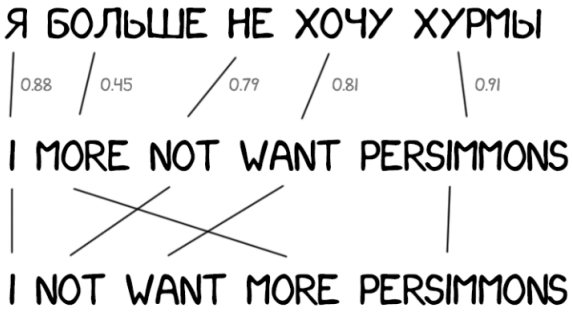
Для определения близости предложений можно использовать tf-idf, Коэффициент Жаккара и другие.

Однословный статистический перевод

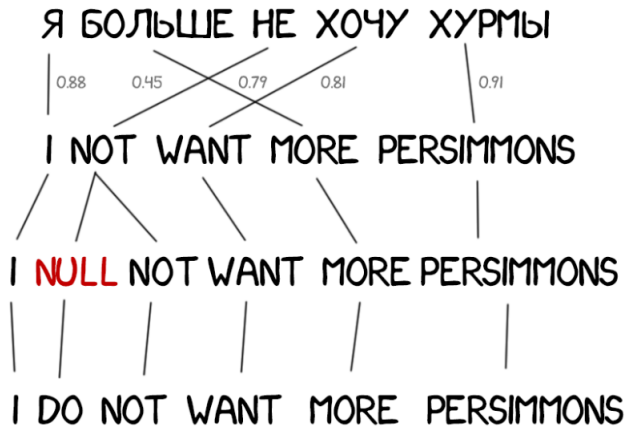
- Мешок слов



- Учет порядка слов в предложении

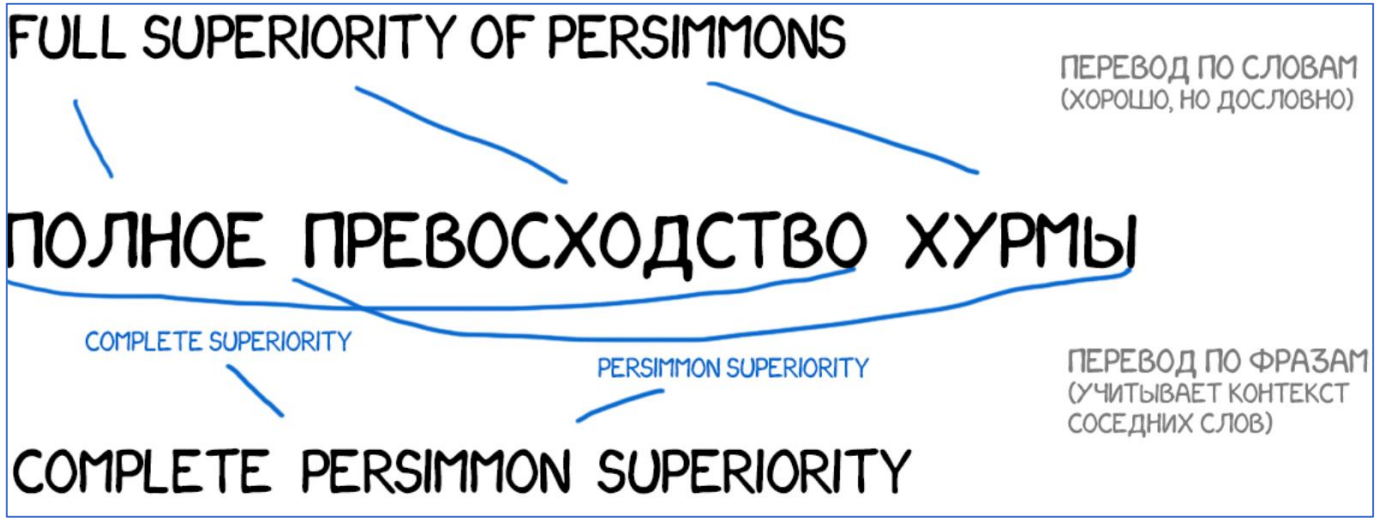


- Добавление отсутствующих слов



- Перестановки слов

Фразовый статистический перевод



Сравнение донейросетевых методов

Перевод на основе правил:

- Стабильный и предсказуемый результат
- Высокая синтаксическая точность
- Долго и сложно обучать модель
- Неспособность адаптироваться к новым данным

Статистический перевод:

- Более точный
- Не нужны лингвисты
- Статистические аномалии
- Простота настройки обучения модели

Нейросетевой подход

Метрики оценки качества машинного перевода

- Автоматическая оценка
  1. BLEU
  2. Расстояние Левенштейна
  3. METEOR

- Человеческая оценка
  1. ALPAC
  2. ARPA

Bilingual evaluation understudy

Cand 1: Mary no slap the witch green
Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Точность кандидата 1 по 1-граммам составляет 5/6

Bilingual evaluation understudy

Cand 1: Mary no slap the witch green.
Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Точность кандидата 1 по 2-граммам составляет 1/5

Bilingual evaluation understudy

Cand 1: Mary no slap the witch green.
Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Точность кандидата 2 по 1-граммам составляет 7/10

Bilingual evaluation understudy

Cand 1: Mary no slap the witch green.
Cand 2: Mary did not give a smack to a green witch.

Ref 1: Mary did not slap the green witch.
Ref 2: Mary did not smack the green witch.
Ref 3: Mary did not hit a green sorceress.

Точность кандидата 2 по 2-граммам составляет 4/9

Bilingual evaluation understudy

Cand 1: $p = \sqrt[2]{\frac{5}{6}\frac{1}{5}} = 0.408$

Cand 2: $p = \sqrt[2]{\frac{7}{10}\frac{4}{9}} = 0.558$

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \le r \end{cases}$$

В итоге: BLUE = BP x p

Пусть r - длина экспертного предложения с наибольшим количеством совпадающих N-грамм. Пусть c - длина машинного перевода

Cand 1: Mary no slap the witch green.

Best Ref: Mary did not slap the green witch.

$c = 6, \quad r = 7, \quad BP = e^{(1-7/6)} = 0.846$

$BLEU = 0.846 \times 0.408 = 0.345$

Cand 2: Mary did not give a smack to a green witch.

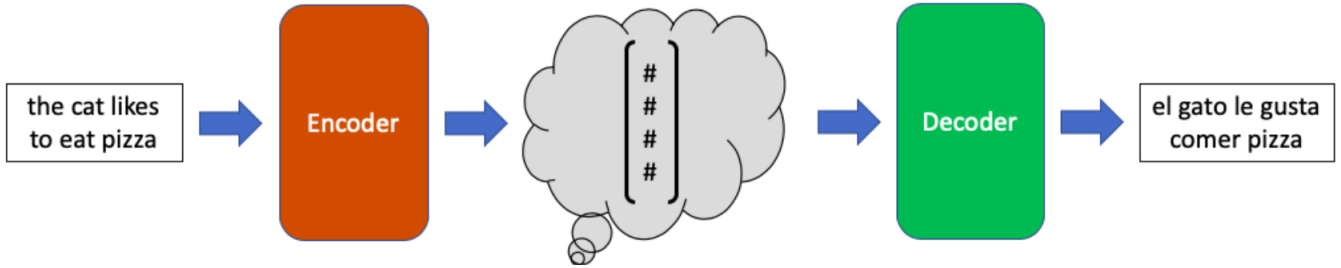
Best Ref: Mary did not smack the green witch.

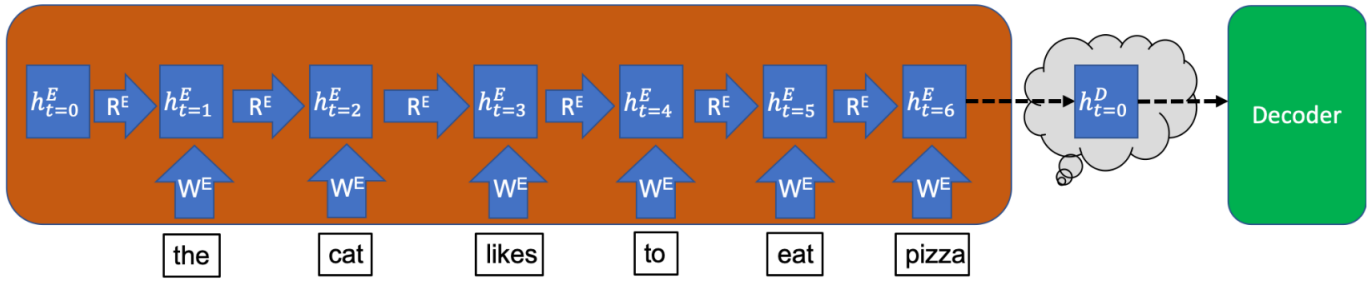
$c = 10, \quad r = 7, \quad BP = 1$

$BLEU = 1 \times 0.558 = 0.558$

Sequence to sequence

- Для начала преобразовываем наши текстовые данные в числовую форму (например, с помощью Embedding методов)
- Далее работаем со структурой Encoder-Decoder

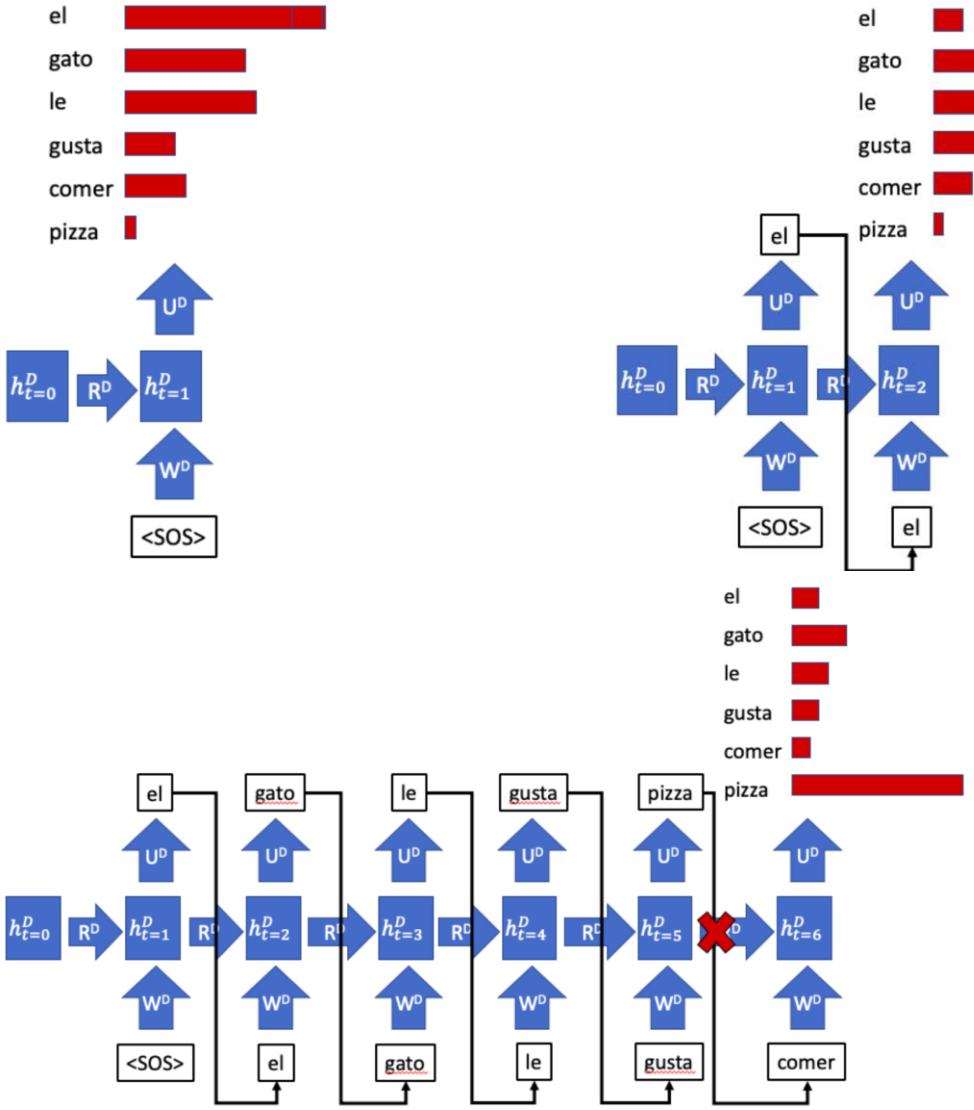




На приведенном выше рисунке синие стрелки соответствуют весовым матрицам, которые мы будем улучшать с помощью обучения для достижения более точных переводов

Sequence to sequence

Последовательность шагов декодера



Sequence to sequence

Cross-Entropy Loss

$$-\sum_{w=1}^{|S|}\sum_{e=1}^{|V|} y_{w,e}\log(\hat{y}_{w,e})$$

$|S|$ = Length of Sentence

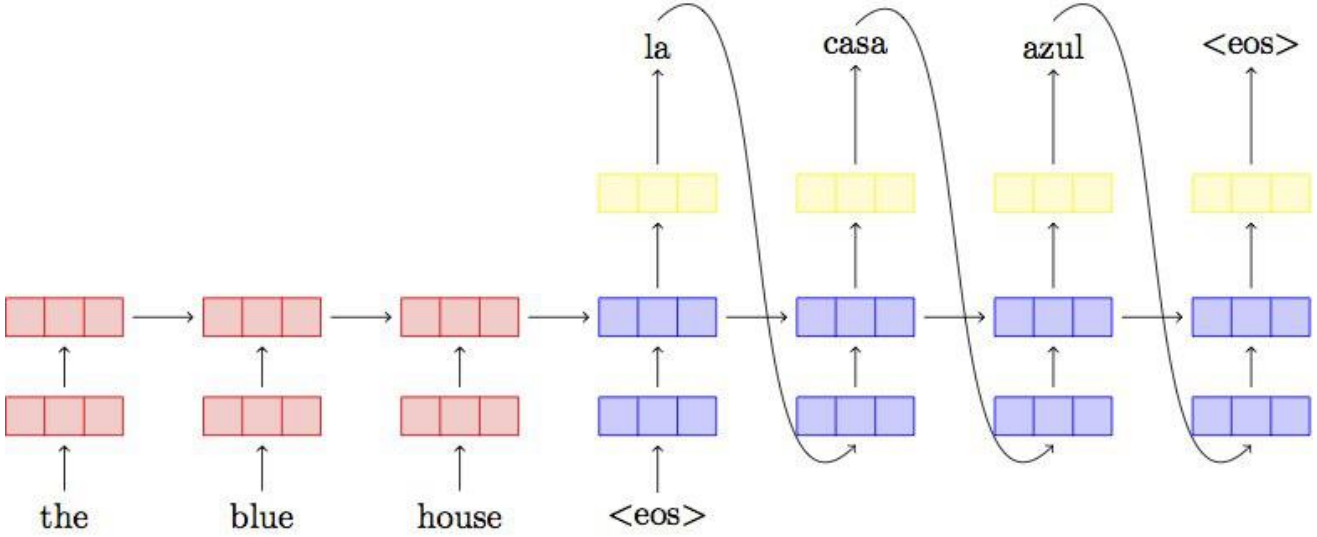
$|V|$ = Length of Vocabulary

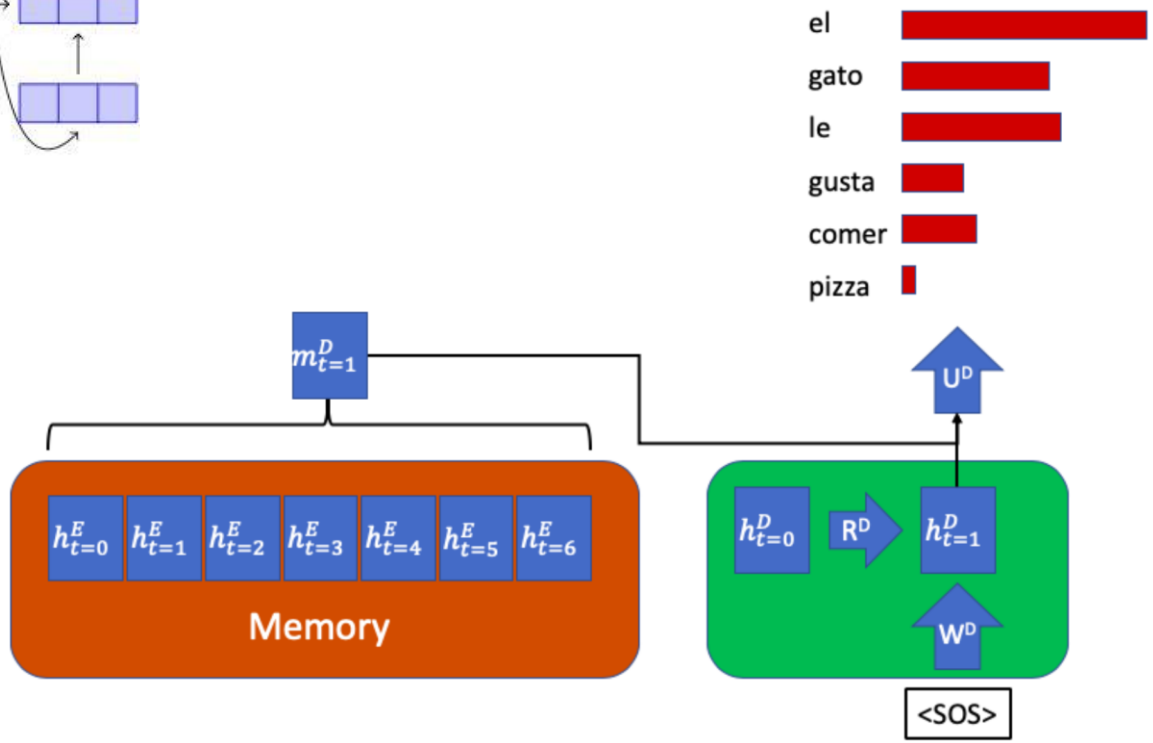
$\hat{y}_{w,e}$ = predicted probability of vocab entry e on word w.

$y_{w,e}$= 1 when the vocabulary entry is the correct word

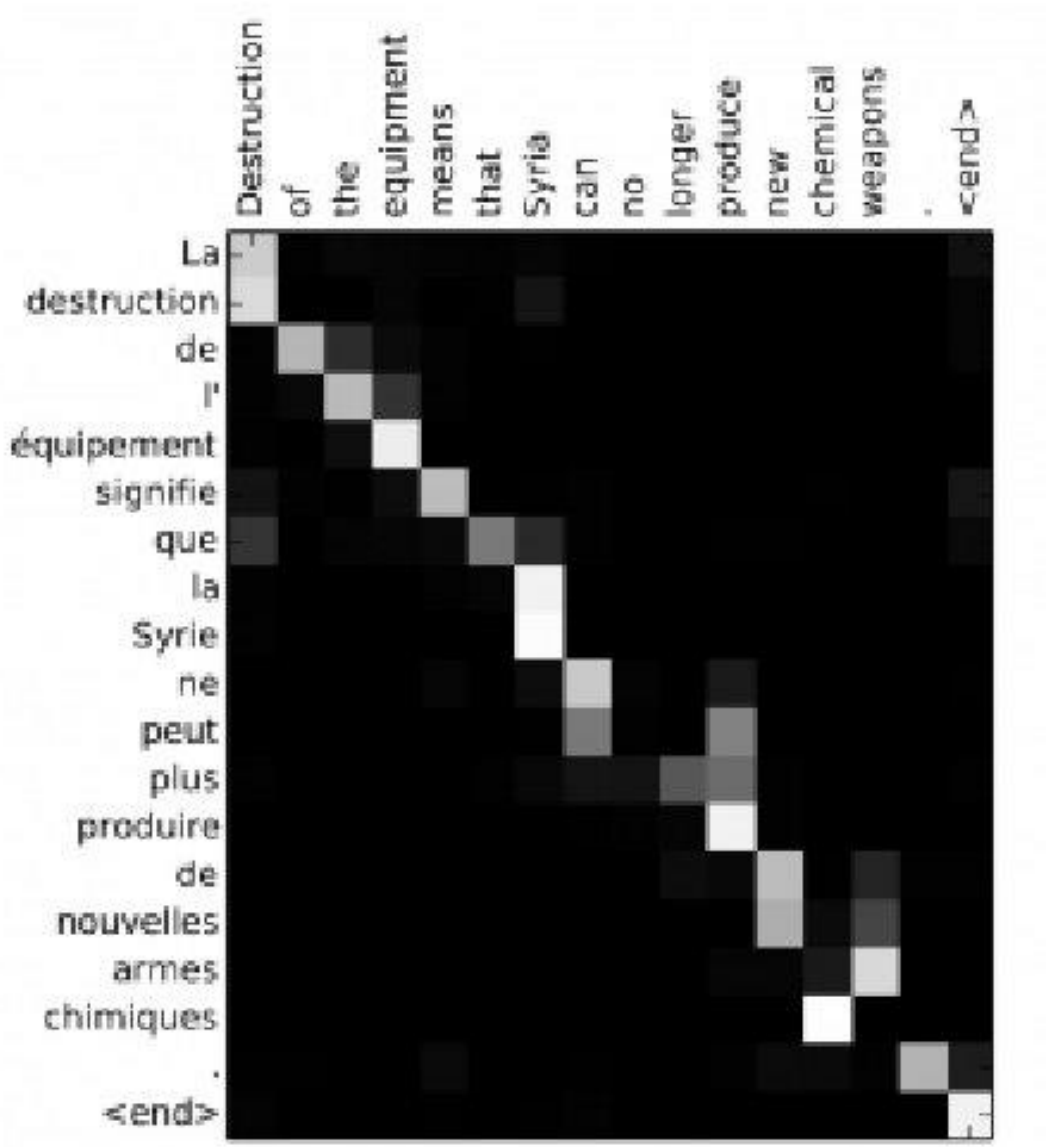
$y_{w,e}$= 0 when the vocabulary entry is not the correct word

Sequence to sequence with attention





Sequence to sequence with attention



Сравнение нейронного подхода со статистическим

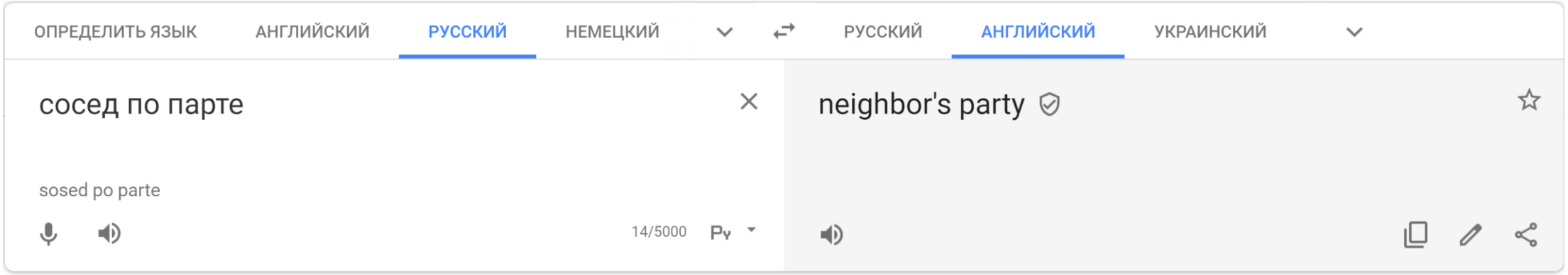
Results:

System ↓	Law	Medical	IT	Koran	Subtitles
All Data	30.5 32.8	45.1 42.2	35.3 44.7	17.9 17.9	26.4 20.8
Law	31.1 34.4	12.1 18.2	3.5 6.9	1.3 2.2	2.8 6.0
Medical	3.9 10.2	39.4 43.5	2.0 8.5	0.6 2.0	1.4 5.8
IT	1.9 3.7	6.5 5.3	42.1 39.8	1.8 1.6	3.9 4.7
Koran	0.4 1.8	0.0 2.1	0.0 2.3	15.9 18.8	1.0 5.5
Subtitles	7.0 9.9	9.3 17.8	9.2 13.6	9.0 8.4	25.9 22.1

Сравнение нейронного подхода со статистическим

	SMT	NMT
Core element	Words	Vectors
Knowledge	Phrase table	Learned weights
Training	Slow Complex pipeline	Slower More elegant pipeline
Model size	Large	Smaller
Interpretability	Medium	Very low Opaque translation process
Introducing ling. knowledge	Doable	Doable (yet to be done!)
Open source toolkit	Yes (Moses)	Yes (many!)
Industrial deployment	Yes	Yes (now at google, systran, wipo)

Заключение



Вопросы для самостоятельной

1. Опишите основную концепцию (последовательность) работы алгоритма в статистической модели перевода
2. В чем основные преимущества статистических моделей по сравнению с переводами на основе правил?
3. Какие главные проблемы нейронного машинного перевода?

Ссылки на источники

- https://arxiv.org/pdf/1508.04025.pdf
- https://towardsdatascience.com/neural-machine-translation-15ecf6b0b
- http://tpc.at.ispras.ru/wp-content/uploads/2011/10/lecture9-2012.pdf
- https://arxiv.org/pdf/1406.1078.pdf
- http://vas3k.ru/blog/machine_translation
- https://www.aclweb.org/anthology/D13-1176.pdf
- http://lig-membres.imag.fr/blanchon/SitesEns/NLSP/resources/SMT-vs-NMT.pdf