Нейросетевые языковые модели

Евгений Борисов

Нейросетевые языковые модели

Языковая модель
- предсказываем следующее слово на основе предыдущих
- оценка (вероятность) совместимости цепочки слов

Оценка цепочки слов (биграммная модель):

$$p(w_1 \ldots w_n) = \prod_{k=1}^{n} p(w_k | w_{k-1})$$

$$p(w_n | w_{n-1}) = \frac{C(w_{n-1} w_n)}{C(w_{n-1})}$$

P - вероятность совместного использования слов

C(w) – количество слов w в тексте

Приложения

распознавание речи

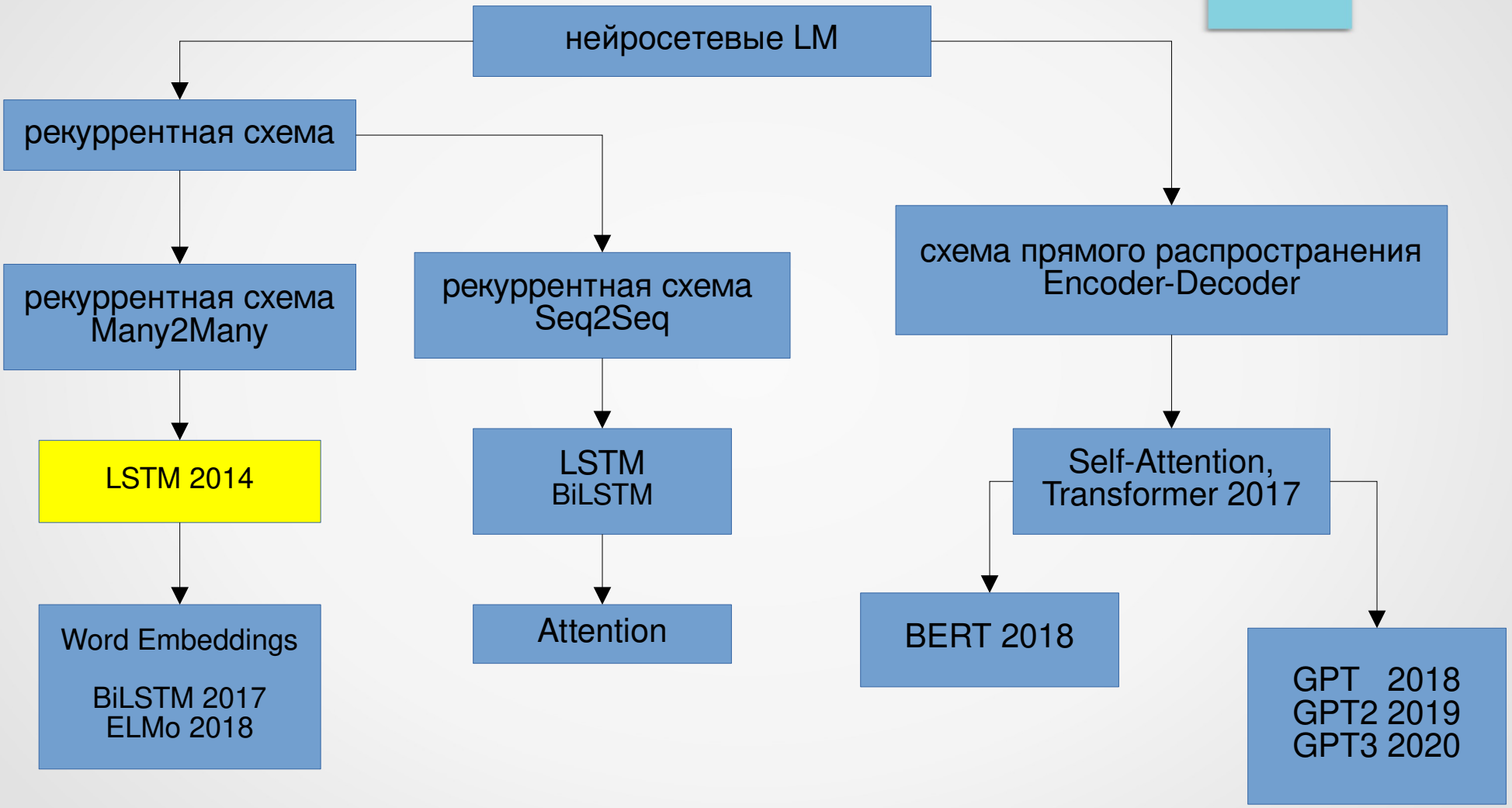
определение частей речи

генерация текстов

извлечение терминов

поиск и коррекция семантических ошибок

Нейросетевые языковые модели

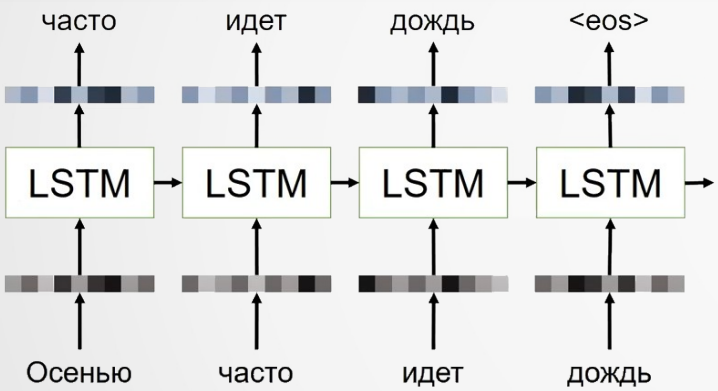


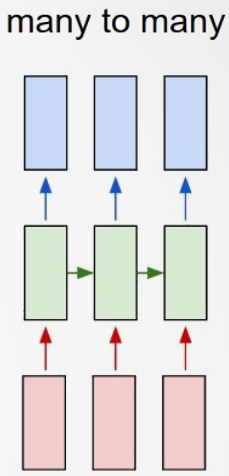
Нейросетевые языковые модели

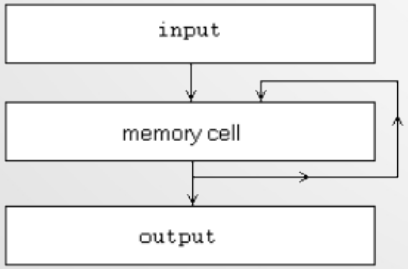
Простая схема с рекуррентной сетью

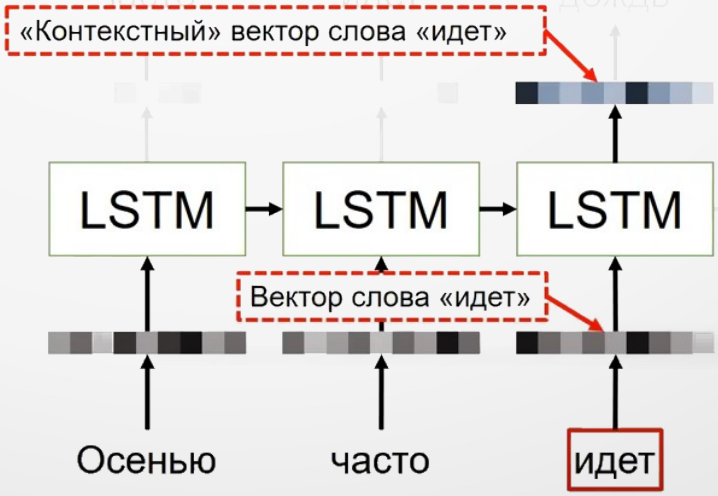
Предсказываем следующее слово по предыдущему контексту

получаем word embedding, который учитывает левый контекст



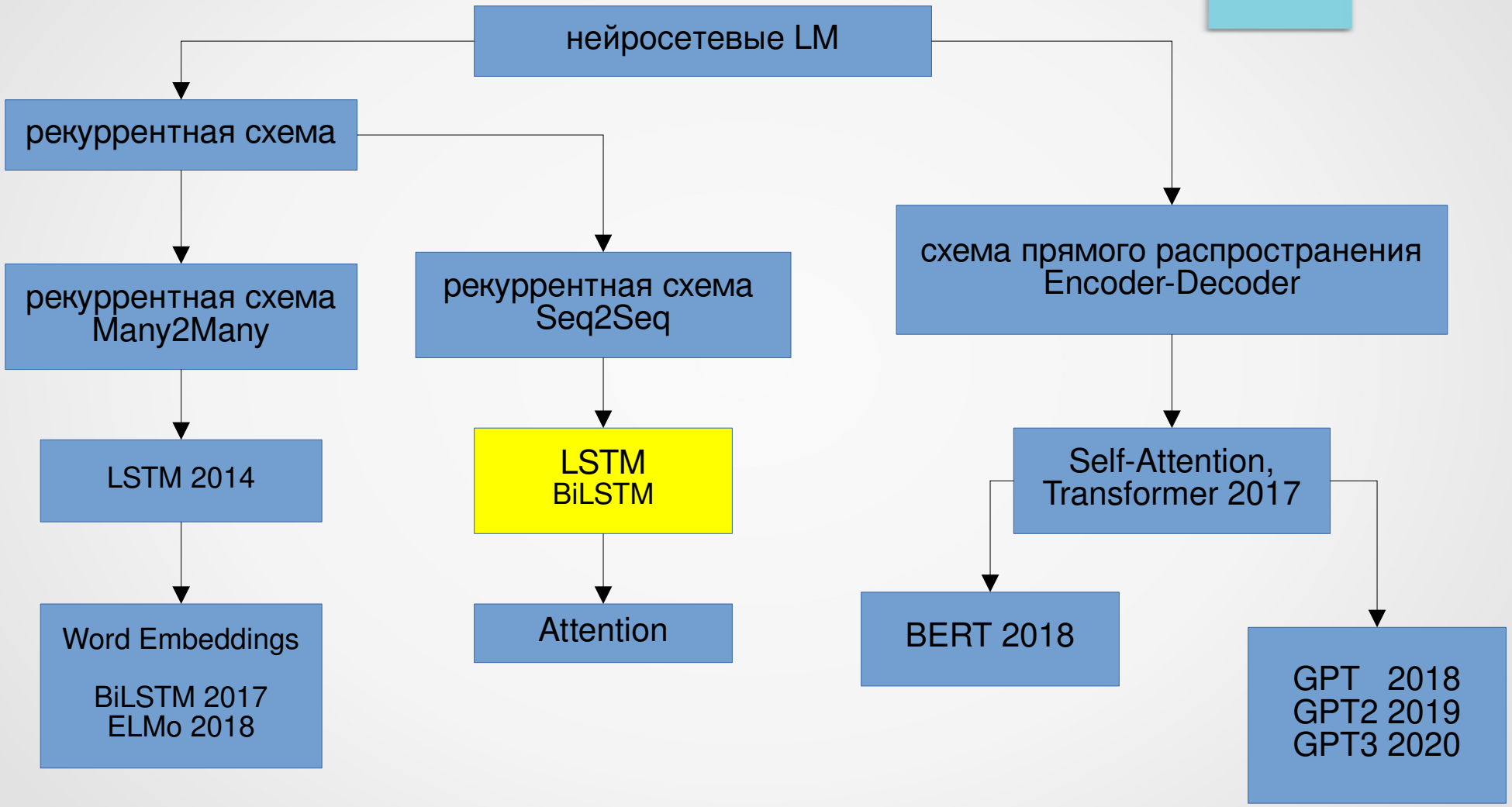






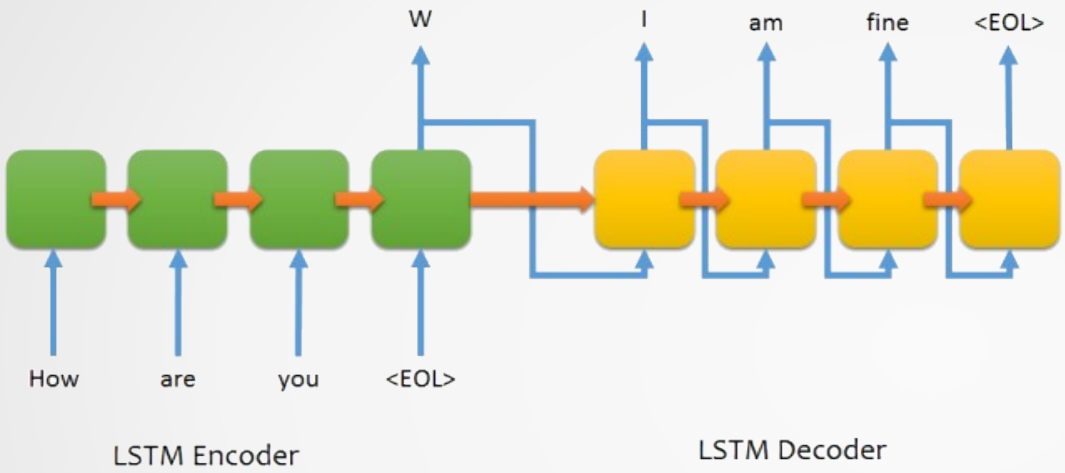
https://www.youtube.com/watch?v=_8MGdpt4I9M

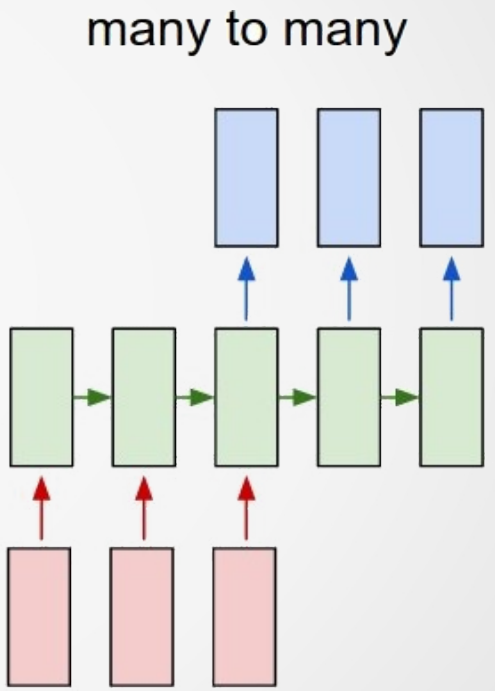
Нейросетевые языковые модели



Нейросетевые языковые модели

Рекуррентная схема SEQ2SEQ





пара рекуррентных неросетей

- кодировщик,
формирует внутреннее представление

- декодировщик,
авторегрессионная модель,
разворачивает состояние энкодера

Нейросетевые языковые модели

git clone *https://github.com/mechanoid5/ml_nlp.git*

Нейчев Радослав Прикладное машинное обучение 3. Machine translation. Лекторий ФПМИ, 2020
https://www.youtube.com/watch?v=6HibilFua-U

Турдаков Д.Ю. Основы обработки текстов. ИСП РАН, 2017,2021

https://www.youtube.com/playlist?list=PL5cBzMoPJgCUn6TbfhqilyToW5lScOdd3

https://www.youtube.com/playlist?list=PL5cBzMoPJgCXFdSvWaunOy4cILirW1lMD