Fine-tuning neural conversation models for auxilary goals by means of deep reinforcement learning

Дмитрий Андреевич Персиянов

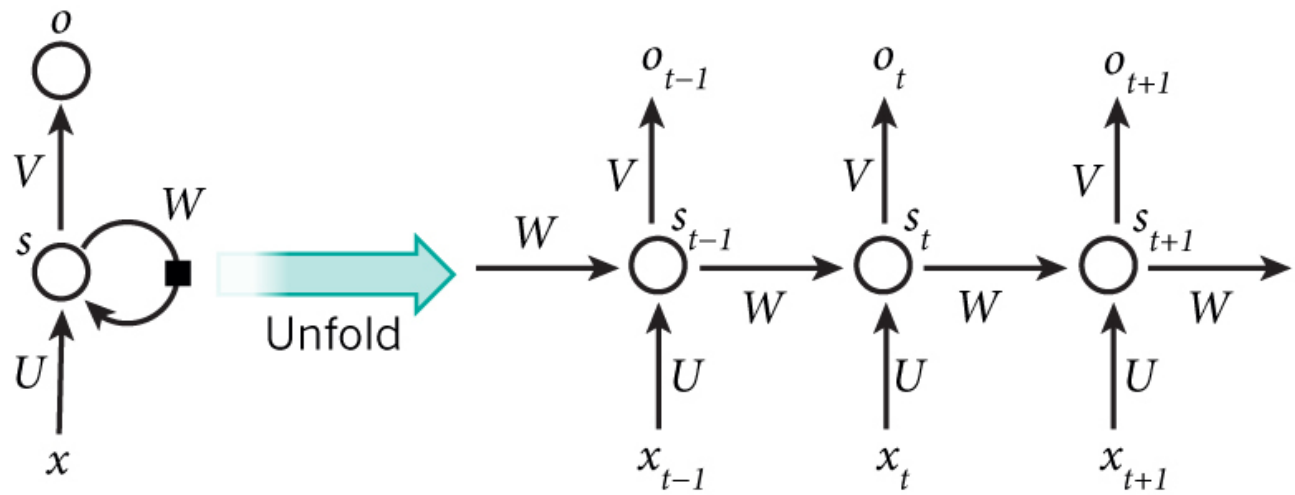
Московский физико-технический институт

22 ноября 2016 г.

План

- Языковые модели и диалоговые системы
- Их проблемы
- Обучение с подкреплением
- Как сравнивать две диалоговые системы

В последнее время рекуррентные сети успешно используется для построения языковых моделей. Обучение происходит на огромных корпусах текстов.



RNN являются языковыми моделями, потому что выучивают распределение $p(w_t|w_{t-1}, \cdots, w_1)$.
Обучаются, минимизируя кроссентропию:

$$J = -\frac{1}{T}\sum_{t=1}^{T}\sum_{j=1}^{|V|} y_{t,j} \cdot log(\hat{y}_{t,j}),$$

где y_t – one-hot вектор длиной $|V|$, кодирующий t-ое слово в предложении, $\hat{y}_t$ – распределение вероятностей следующего слова, полученное из нейросети.

A Neural Conversational Model

Одна из первых попыток обучить RNN отвечать на сообщения. (https://arxiv.org/pdf/1506.05869.pdf)

Describe your problem: browser keeps crashing
Machine: *hi*
Human: *hi*
Machine: *could you please let me know what is the operating system you are using ?*
Human: *linux*
Machine: *ok this is the issue with the system performance , nasser , tools .*
Human: *what is that ?*

Human: *what is your job ?*
Machine: *i 'm a lawyer .*
Human: *what do you do ?*
Machine: *i 'm a doctor .*

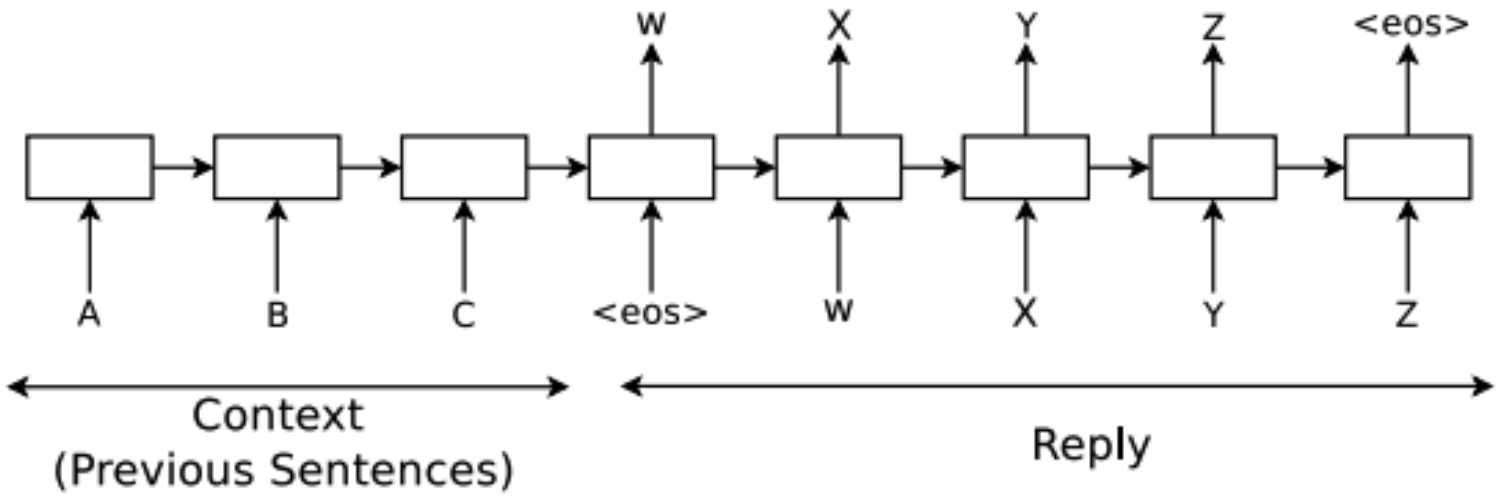


Figure 1. Using the *seq2seq* framework for modeling conversations.

Диалоговые системы

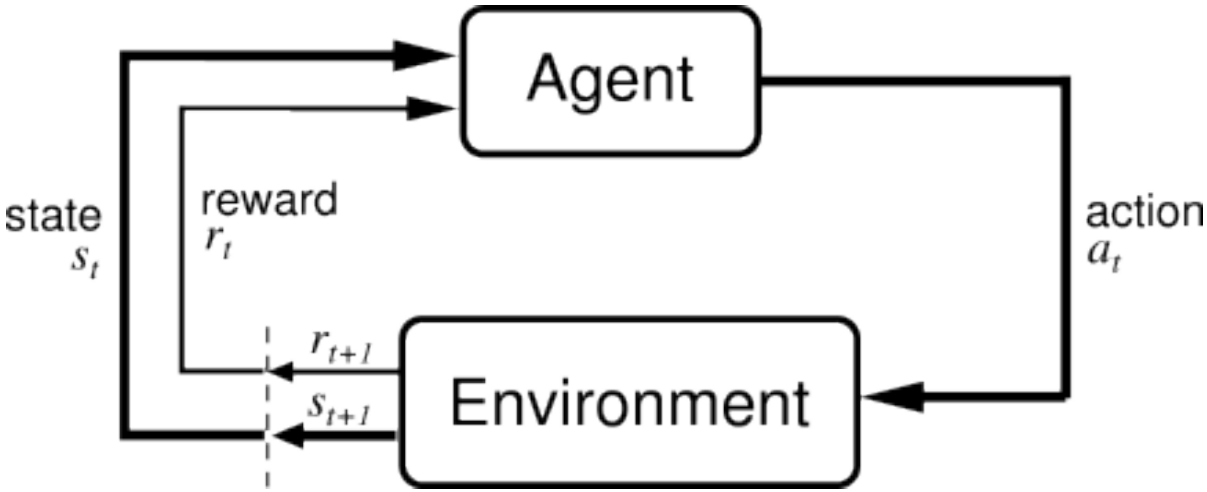
Дальнейшее развитие:

- Механизм внимания в RNN
- Bidirectional RNNs
- Hierarchical models
- Persona-based seq2seq

Диалоговые системы, проблемы

- На один и тот же вопрос два разных ответа (inconsistency)
- Нейронная сеть выучивает языковую модель минимизируя кроссентропию, а нам иногда хочется другого:
  - Консистентность (учитывание контекста предыдущих ответов)
  - Не использование каких-то слов (табу)
  - **Ведение беседы в каком-то стиле (говорить как Путин)**
  - **Максимизация скорости завершения диалога**
  - **Максимизация удовлетворенности пользователя**
  - Максимизация ...

Обучение с подкреплением



Необходимо найти стратегию $\pi(a|s)$, такую что

$$E_\pi \left[R_0 + \gamma R_1 + \cdots + \gamma^t R_t + \cdots \right] \to max.$$

Обучение с подкреплением

В диалоговых системах:

- Действия a – слова (предложения), которые мы генерируем
- Стратегия π – распределение, которое выучивает нейронная сеть
- Награды R – задаются по-разному, в зависимости от задачи

Обучение с подкреплением

Related papers:

- A Network-based End-to-End Trainable Task-oriented Dialogue System (https://arxiv.org/pdf/1604.04562v2.pdf)
- Deep Reinforcement Learning for Dialogue Generation (https://arxiv.org/pdf/1606.01541v4.pdf)
- Semantically Conditioned LSTM-based Natural Language Generation for Spoken Dialogue Systems (https://arxiv.org/pdf/1508.01745v2.pdf)

Метрики для сравнения двух моделей

- Оценка правдоподобия (или перплексии) моделей на валидационной выборке
- n-gram based метрики: BLEU, WER, METEOR
- Оценка ответов двух моделей асессорами
- Обучение дискриминаторов для двух моделей

План исследования

- Уже есть baseline neural conversational model.
- Применение Policy Gradient методов для дообучения модели под разные задачи.
- Problem 1. Научится говорить как конкретный человек, etc.
    - **Baseline**: Finetuning по классическому LLH лоссу.
    - **Hypothesis**: Baseline получиться побить, если дообучать RL лоссом.
- Problem 2. Максимизировать удовлетворенность пользователя ответом, etc.
    - **Hypothesis**: RL in continuous action spaces. Подмена вектора енкодера другим, сгенерированным с помощью стратегии $\pi(a|s)$.