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

Дмитрий Андреевич Персиянов

Московский физико-технический институт

22 июня 2017 г.

План

- Conversational модели
- RL дообучение
- BePolite эксперимент
- BeLikeX эксперимент
- Заключение и дальнейшие исследования

Conversational модели

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

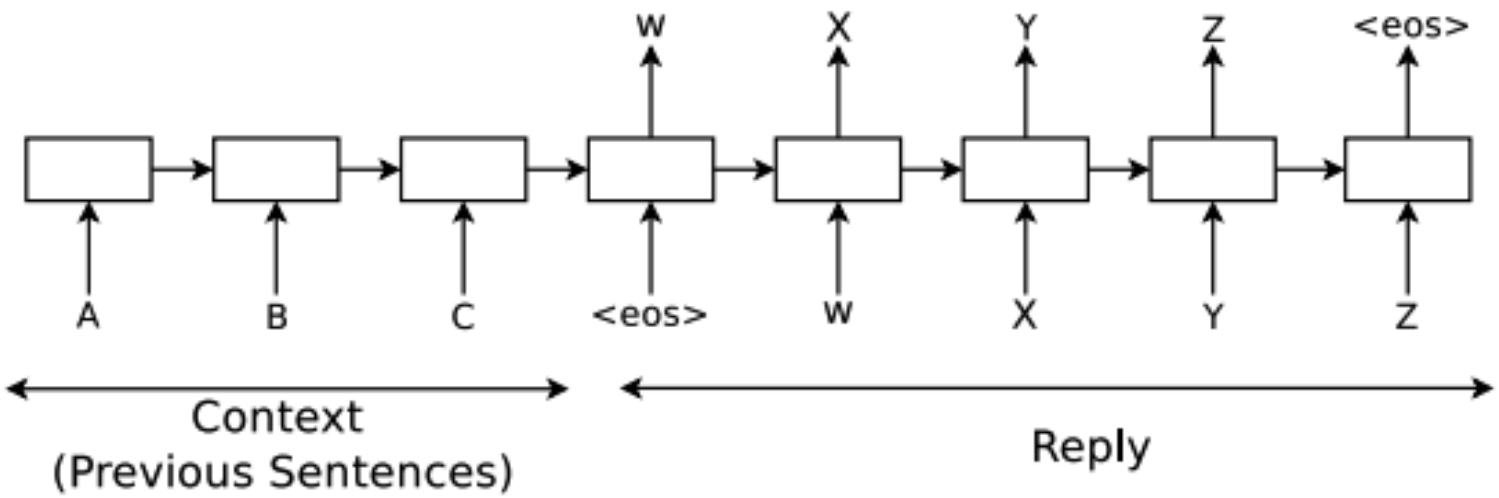


Figure 1. Using the *seq2seq* framework for modeling conversations.

Имея обучающий пример $(\mathbf{c}, \mathbf{a})$ контекст-ответ, где $\mathbf{c} = \{c_1, c_2, \ldots, c_n\}$, $\mathbf{a} = \{a_1, a_2, \ldots, a_k\}$, учим модель, минимизируя лосс:

$$L(\theta) = -\sum_{t=1}^{k} \log \left(p_\theta(a_t | a_1, \ldots, a_{t-1}, \mathbf{c}) \right)$$

или (в RL нотации)

$$L(\theta) = -\mathbb{E}_{\mathbf{c},\mathbf{a} \sim \mathcal{D}} \left[\log p_\theta(\mathbf{a}|\mathbf{c}) \right]$$

Conversational модели, проблемы

- На один и тот же вопрос два разных ответа (inconsistency)
- Выучиваем, минимизируя кроссентропию, а нам иногда хочется другого:
  - Консистентность (учитывание контекста предыдущих ответов)
  - **Запрет на использование каких-то слов**
  - **Ведение беседы в каком-то стиле**
  - Максимизация скорости завершения диалога
  - Максимизация удовлетворенности пользователя
  - Максимизация ...

RL дообучение

Диалоговую модель $p_\theta(a_t|h_t, a_{t-1})$ можно воспринимать как политику $\pi_\theta(a_t|s_t)$.
Необходимо найти политику $\pi(a|s)$, такую что

$$\mathbb{E}_{\hat{\mathbf{a}}\sim\pi}\left[R_0 + \gamma R_1 + \cdots + \gamma^t R_t + \cdots\right] \to max,$$

где $R(\mathbf{a}, \hat{\mathbf{a}})$– некоторая функция награды, зависящая от правильного ответа $\mathbf{a}$ из обучающей выборки и сгенерированного моделью ответа $\hat{\mathbf{a}}$.
Также возможен более гранулярный вариант $R(a_t, \hat{a}_t)$.

BePolite эксперимент

- Данные: opensubtitles.org (en), 18млн пар (контекст, ответ).
- Собрали 800 обсценных слов (маты, религиозные/расовые оскорбления). Обозначим это множество за $\mathcal{S}$.
- Функция наград: $R(\hat{a}_t) = -\mathbb{I}[\hat{a}_t \in \mathcal{S}]$
- Используем предобученную по MLE лоссу модель.
- Дообучаем policy-gradient методом по $L(\theta) = -\mathbb{E}_{\hat{\mathbf{a}} \sim p_\theta}\left[\sum_{t=1}^{k} R(\hat{a}_t) \log p_\theta(\hat{a}_t | \hat{a}_{t-1}, \dots)\right] - \alpha \mathbb{E}_{\mathbf{a} \sim \mathcal{D}}\left[\log p_\theta(\mathbf{a})\right]$
- $\alpha = 5, 20$.
- Обучаем 500 батчей по 64 примера.

BePolite эксперимент

Таблица: Метрики бейзлайна

Средняя награда	Перплексия
-0.136	3.142

Таблица: Метрики после policy-gradient дообучения

α	Средняя награда	Перплексия
5	-0.021	3.297
20	-0.065	3.270

BeLikeX эксперимент

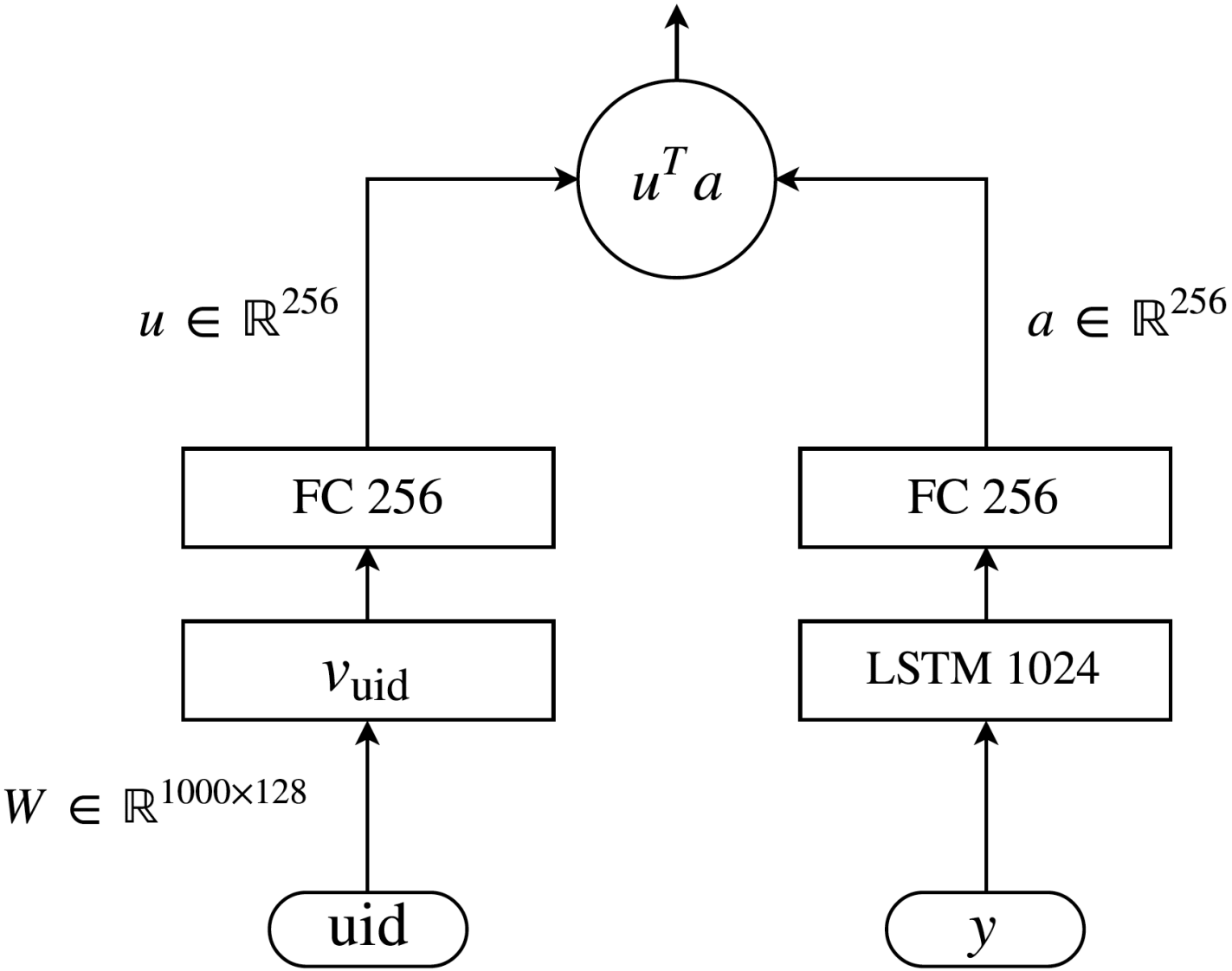
- Данные $\mathcal{D}_g$: twitter (ru), 50 млн примеров (контекст, ответ) для бейзлайна.
- Данные: twitter (ru), 1.2 млн примеров с идентификаторами пользователей.
- Отобрали 1000 пользователей по частоте участия в диалогах. (Топ1 – 9500 ответов на чьи-то твиты). Обозначим $\mathcal{D}_u$.
- Обучили dssm-like модель $D(\mathbf{uid}, \mathbf{a}) \in [-1, 1]$ в качестве прокси-награды.
- Выбрали одного юзера – тех. поддержка. Обозначим соотв. выборку $\mathcal{D}_s$.

BeLikeX эксперимент: примеры фраз

Таблица: Примеры фраз тех. подержки

Контекст	Ответ
Жаль, что с я связан корпоративным тарифом. Очень хочется перейти на оператора с нормальным покрытием. **EOS** и мегафон здесь не берёт, и папа джонс не везёт пиццу **EOS**	**BOS** Добрый вечер! Укажите, пожалуйста, точный адрес (нас. пункт, улицу, номер дома), в чем именно заключается сложность. Проверим. **EOS**
на модем который используется как роутер списаны звонки, и всякое такое чего у меня в тарифе нету... **EOS** сервис на высшем уровне! Такое сейчас время, не своруешь не проживёшь , жаль ! Хорошая была компания! **EOS**	**BOS** коллеги, обратите внимание. **EOS**
Добрый вечер! С какими сложностями Вы столкнулись, скажите, пожалуйста? Постараемся Вам помочь. **EOS** интернет работал только "Е"полтора часа,уже все хорошо. **EOS**	**BOS** Рады, что Ваш вопрос решился. Возникнут вопросы - обращайтесь! =) **EOS**

BeLikeX эксперимент: DSSM



BeLikeX эксперимент: DSSM

$$D_\psi(\text{uid}, y) = \frac{u^T a}{\|u\|\|a\|}, \tag{1}$$

$$L(\psi) = \mathbb{E}_{\text{uid}, y_{\text{pos}}, y_{\text{neg}} \sim \mathcal{D}_u} \left(\max(0, 0.5 - D_\psi(\text{uid}, y_{\text{pos}}) + D_\psi(\text{uid}, y_{\text{neg}})) \right) \tag{2}$$

BeLikeX эксперимент: DSSM ranking

Таблица: Фразы других юзеров, которым DSSM выдал награду 1.0

Фраза
Ответили в ДМ.
Приносим извинения за возможные неудобства.
сейчас все решим, приносим извинения за неудобства.
позвольте еще раз принести Вам наши извинения за доставленные неудобства.
Спасибо, что помогли нам провести работу над ошибками.
Если Вы оставляли заявку на тестирование, появится после праздников.
Ответили вам в ЛС.
Оператор видит ситуацию происходящую с номером и сможет вам помочь.

BeLikeX эксперимент: SCST дообучение

$$\Delta\theta = \nabla_\theta \log p_\theta(\hat{y}|\mathbf{h}_\mathbf{0}^{\text{dec}}) \cdot (D(\hat{y}) - D(\overline{y})), \tag{3}$$

где $\overline{y}$ – жадный ответ, а $\hat{y}$ – просемплированный из политики.

BeLikeX эксперимент: результаты

Таблица: Перплексия.

	$\mathcal{D}_g$	$\mathcal{D}_s$
BASELINE	6.330	14.269
LLH-FINETUNED	24.308	1.040
SCST-ON-SUPPORT	17.574	1.175
SCST-ON-USERS	8.178	26.691
SCST-ON-HIGH-REWARDED	24.305	1.283

BeLikeX эксперимент: результаты

	$\mathcal{D}_g$	$\mathcal{D}_s$
BASELINE	0.0130	0.0484
LLH-FINETUNED	0.6743	0.9546
SCST-ON-SUPPORT	0.7798	0.8967
SCST-ON-USERS	0.0192	0.0183
SCST-ON-HIGH-REWARDED	0.3357	0.9331

Таблица: Значения avgD для моделей из эксперимента BeLikeX.

BeLikeX эксперимент: примеры ответов

Таблица: Примеры ответов моделей.

Контекст	SCST-ON-SUPPORT	SCST-ON-HIGH-REWARDED
привет! **EOS** как дела?	коллеги помогите , пожалуйста , только в личные сообщения (город , улица , поселок , лк)	у нас нет света , но хорошо .
интернет не работает что делать?' **EOS**	коллеги помогите , пожалуйста , более подробно какие сложности с использованием .	здравствуйте ! если возникли сложности , то вы можете проверить баланс ?

BeLikeX эксперимент: примеры ответов

Таблица: Примеры ответов моделей.

что такое любовь ? **EOS**	коллеги, пожалуйста , базовые всей компании	все ахаха-хаххахаха
здравствуйте ! с какими сложностями столкнулись ? **EOS** если будут вопросы , пишите !	если возникнут вопросы , пожалуйста , обращайтесь	если будут вопросы , пишите !
какой у вас регион ? **EOS** томская область	какой у вас регион ?	проверьте , пожа-луйста

BeLikeX эксперимент: примеры ответов

Таблица: Примеры ответов моделей.

какой у вас регион ? **EOS** как дела ?	проверьте , пожа- луйста	:d все хорошо , но ограничения к сведению
ты любишь учиться ? **EOS** я обожаю ! а ты ?	какой у вас регион ?	нет , но к сожалению , не могу сказать одно время , да .

Похожие работы

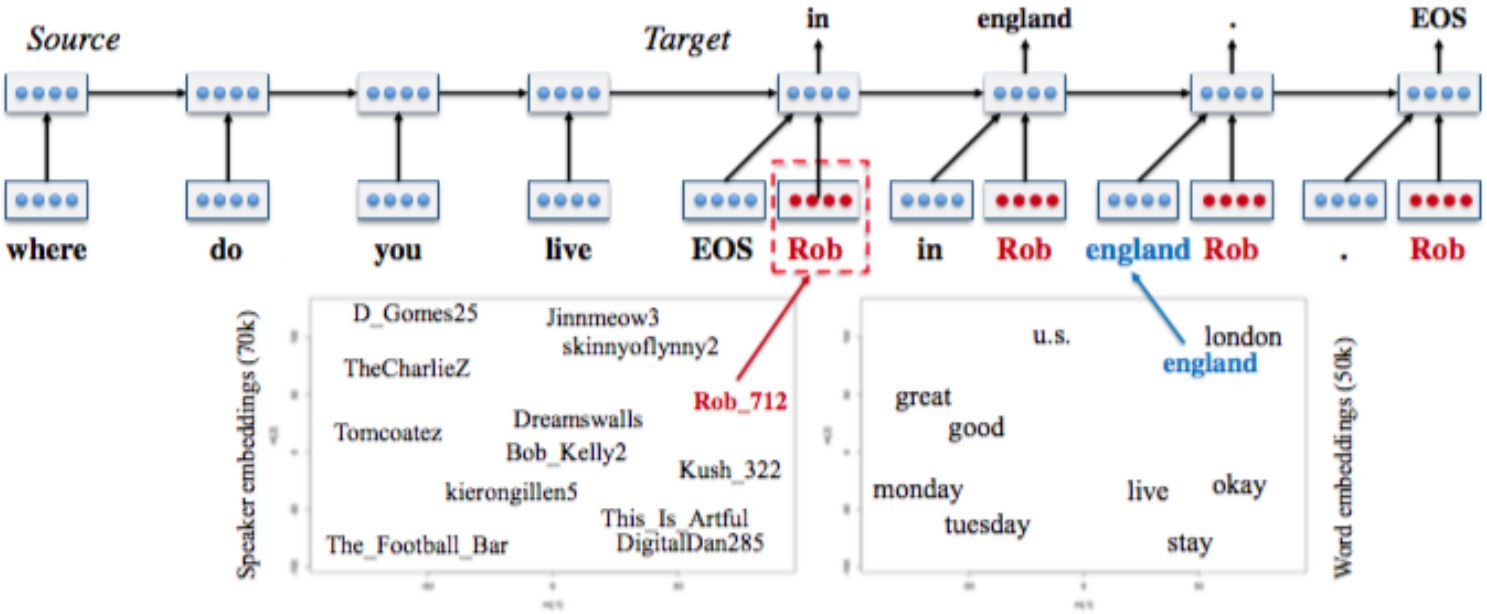
- Deep Reinforcement Learning for Dialogue Generation (https://arxiv.org/pdf/1612.00563.pdf) – дообучают RL-ом, но борются с проблемой затухания диалогов и общих ответов.
- A Persona-Based Neural Conversation Model (https://nlp.stanford.edu/pubs/jiwei2016Persona.pdf) – выучивают ембеддинги для пользователей и подают на вход декодеру.

Похожие работы

- Deep Reinforcement Learning for Dialogue Generation (https://arxiv.org/pdf/1612.00563.pdf) – дообучают RL-ом, но борются с проблемой затухания диалогов и общих ответов.

Похожие работы

- A Persona-Based Neural Conversation Model (https://nlp.stanford.edu/pubs/jiwei2016Persona.pdf) – выучивают ембеддинги для пользователей и подают на вход декодеру.



Заключение и дальнейшие исследования

- RL помогает быстро и эффективно дообучать модели под разные требования, выразимые в виде функции наград.

- BePolite: посмотреть как запрет одних слов влияет на частоту использования семантически близких, но которых нет в словаре
- BeLikeX: использовать дискриминатор, обученный лишь на одном юзере, как в GAN'ах. Пытаться обмануть его.
- BeLikeX: обусловить дискриминатор на контекст, посмотреть что получится.