TFS19s: Машинное обучение в диалоговых системах

Шутейки

Кармазин Василий
Юсов Александр

Описание задачи

“По мотивам поста запилить бота, который шутит шутейки.”

В статье Humor Generation with Recurrent Neural Networks автор использует:

1. Датасет с короткими шутками (50-150 символов)
2. Character level LSTM модель (2-3 слоя, по 1024-2048 нейронов)

Результаты автора поста

Why did the cowboy buy a frog? Because he didn't have any brains.
Why can't Dracula be true? Because there are too many cheetahs.
What do you call a black guy flying a plane? A pilot, you racist!

Поиск данных

- В vk.com нашли группы, в которых постятся исключительно текстовые шутки. Через vk api выкачили их себе

  С таких групп удалось собрать ~7k шуток

- Из Цитатник-а Рунета bash.im смогли выкачать ~73k шуток

  Впоследствии пришлось отказаться от данных из-за своей зашумлённости и отклонения данных от темы проекта

- На сайте anekdot.ru мы спарсили ~74k шуток

Обработка данных

1. Оставили только символы из алфавита и некоторую пунктуацию
2. Почистили данные от ложной информации: картинки, реклама
3. Отсеили слишком длинные шутки, где >300 символов
4. Убрали из текста разный мусор в виде: хештегов, номеров, авторов, ссылки и тд
5. Убрали ненормативную лексику

После обработки данных у нас получился датасет в ~50k шуток

Telegram: @TFS_shutki_bot

TFS_shutki_bot

Ахахах) 🤡

Сейчас у нас запущен бот в телеграме, ему можно написать :)

Команды /start или /help вызывают справку

Шутки из второй модели не работают по техническим причинам

Эксперимент 1. LSTM/RNN, Seq2Seq

1. Обучили языковую модель на данных, которые удалось найти в vk
2. Модель:
    a. Использовали двухслойную LSTM со скрытым слоем в 256
    b. Для букв используются эмбединги размерности 128
    c. Предсказания делаются на основе Beam Search с шириной поиска в 5
3. Для генерации шуток подаём модели рандомную строку или рандомный контекстный вектор

Telegram: @TFS_shutki_bot

Результаты LSTM/RNN, Seq2Seq

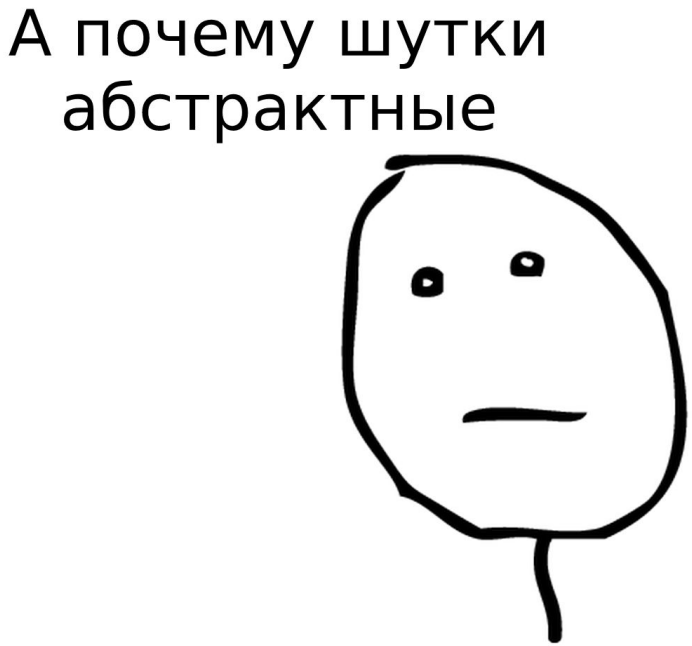
1. интересно, чтобы всегда становится в российских российских российских российских российских россии.

2. интересно, что все равно встретились в россии. народный.

3. анекдоты юмор смех ржака поздравляются детей.

4. российские слова: - это когда- нибудь настроение.

5. интересно, что в россии - это когда- нибудь в россии.



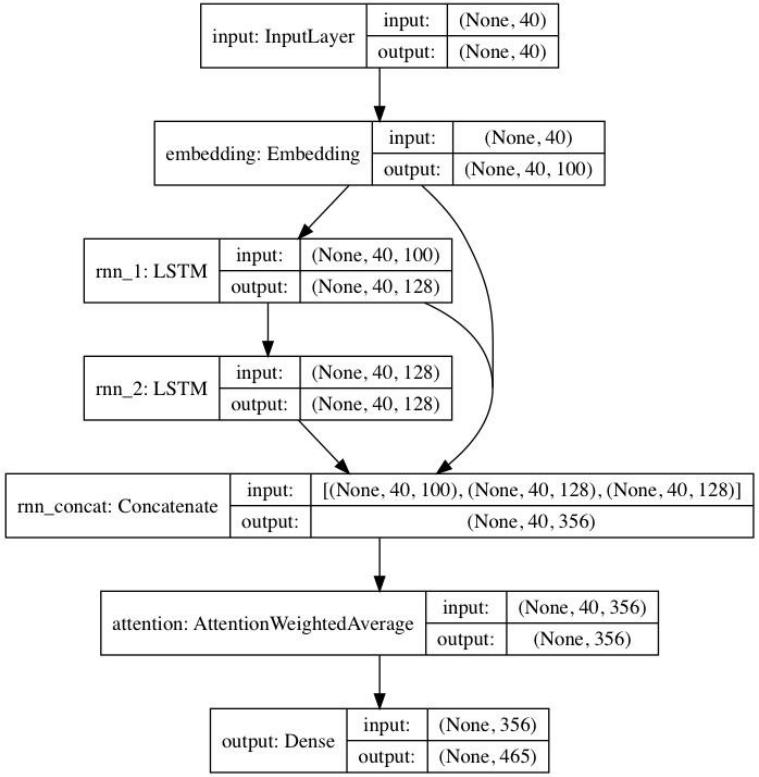
Telegram: @TFS_shutki_bot

Эксперимент 2. LSTM with Attention

Попытка использовать модель построенную на символах с механизмом Attention

1. Взята готовая реализация из textgenrnn, с rnn_layers=4, layer_size=128, dim_embeddings=100,
2. Длительность обучения - 15 эпох, ~ 3 часа в colab’e

Детали архитектуры можно найти в репозитории Max Woolf



Telegram: @TFS_shutki_bot

Результаты LSTM with Attention

1. *Почему ковбой купил лягушку?* - Да, ты же сам выбрала (temp=0.5)

2. *Почему ковбой купил лягушку?* - Нет, я просто подумал, что ты можешь сказать о том, что я его подарил (temp=0.3)

3. *Ковбой купил лягушку, чтобы* смотреть на дом после возврата на меня 2013 года в постели! (temp=0.6)

4. *Ковбой купил лягушку, чтобы* показать наши проблемы (temp=0.5)

5. *Дракула не может жить потому, что* она вышла замуж за компьютером (temp=0.6)

6. *Дракула не может жить потому, что* она была всегда и обезьяна.

Telegram: @TFS_shutki_bot

Эксперимент 3. ULMFiT AWD-LSTM

Попробовали использовать уже обученную на русских текстах языковую модель и затрансферить её на наш датасет

1. Использовали библиотеки fastai и PyTorch
2. Предобученная модель, ULMFit на Тайге - универсальная языковая модель AWD-LSTM
3. Получили word based модель
4. Длительность обучения - 20 эпох, ~ 3 часов в colab'е

Особенности AWD-LSTM (ASGD Weight-Dropped LSTM)

- DropConnect
- Average-SGD (NT-ASGD)

Остальные подробности модели можно найти в статье

Telegram: @TFS_shutki_bot

Результаты ULMFiT AWD-LSTM

1. - Почему украина непобедима грибами да потому что у нее огромная сила в карьере !

2. Новости на тв заблокированы на 10 минут делают смешными пушистыми ким чен заметил : медведева назвали геем

3. Одна прикольная умная женщина разных размеров всегда приносит своему администратору взятку .

4. - Жора , что ты записался на бокс
   - Сарочка , а шо случилось таки спокойной ночи

5. В одесском автобусе :
   - Мужчина ! принесите мне вешалку для одежды !
   Продавщица :
   - Тысяча баксов на проезд .
   - Это таки не подарок !
   - Возьмите , пожалуйста и кошелек . скидка - 50 1500 долларов .

Telegram: @TFS_shutki_bot

Сравнение с шутками из статьи. При генерации с начальными словами:

Ковбой купил лягушку чтобы снижать цену за подбор пельмени .

Ковбой купил лягушку чтобы не потерять своего кота . Он : - Вот , например , я не могу найти кота . Он : - Ну , как же , я же не хочу , чтобы я был в Москве .

Почему ковбой купил лягушку? Потому что у него есть хвост .

Почему Дракула не может существовать? Потому что у него нет денег .

Why did the cowboy buy a frog? Because he didn't have any brains.

Why can't Dracula be true? Because there are too many cheetahs.

What do you call a black guy flying a plane? A pilot, you racist!

Telegram: @TFS_shutki_bot

Что мы не сделали, а надо было...

1. Нормальная валидация: мы не делали графики сходимости функции потерь на eval и train выборках, смотрели на сами значения
2. ELMo эмбеддинги, например, взять у iPavlov
3. Собрать больше данных (раз в 5)
4. Попробовать BPE

Telegram: @TFS_shutki_bot