


Написание анекдотов нейросетью



Семен Кудрявцев
Юлия Пыжак
Елена Цейтина

Мотивация/актуальность работы

Многие работы были посвящены генерации шуток, но довольно малое количество было посвящено генерации именно анекдотов, которые на самом деле сильно отличаются от шуток. Если шутка обычно является неординарным ответом на какой-либо вопрос, то анекдоты представляют собой маленькие истории - с сюжетом, иногда известными персонажами, диалогами. В английской Википедии статья про анекдоты называется Russian jokes.

Команда, роли и задачи каждого участника

На этом этапе...

Семен занимается планированием в трелло и менеджерит проект.

Алена собирает данные (уже собрала!)

Юля делает презентацию (уже сделала!)

Все ищут статьи, чтобы изучить опыт людей, занимавшихся этой темой (уже нашли!)

[доска](#) в трелло с распределением задач

Данные

Для обучения модели генерации будем использовать сайты с анекдотами, например:

- <https://www.anekdot.ru/tags/>
- <https://anekdotov.net/menufull.html>
- <https://nekdo.ru/>
- <https://veselka.mobi/11jan22/anekdot.html>
- <https://shytok.net/anekdots.html>
- http://gorodok.tv/jokes_archive <https://vse-shutochki.ru/anekdoty>

Для обучения модели классификации возьмем короткие художественные рассказы, в которых нет юмористической составляющей.

Бейзлайн

В качестве бейзлайна будем использовать марковские цепи.

Метрики оценки

Для оценки качества сгенерированного текста (насколько это анекдот) мы будем использовать бинарный классификатор, обученный на анекдотах и рассказах.

Если этот этап наши нейроанекдоты пройдут, то затем мы оценим их качество, предложив людям в гугл-формах оценить, насколько они смешные.

План действий

- Файн-тюнинг гпт-2 на анекдотах
- Если с файн-тьюнингом все будет плохо, можно попробовать дообучить берта
- сравнить две модели выше с бейзлайном и друг с другом, определить наилучшую
- построение классификации анекдот/не анекдот
- сделать опрос для оценки качества нейро-анекдотов

Последующие применения

Как мы знаем, на ютубе очень популярны комедийные шоу, и у нас есть идея своего с искусственным интеллектом. По плану шоу к 4 роботам приходит в гости 5-ый робот и начинает рассказывать историю. 4 робота должны угадать, как она заканчивается, но они просто его перебивают и шутят. Один из 4 роботов будет постоянно говорить, что на этот счет он знает анекдот - и рассказывать анекдот. Основу для этого робота мы и будем создавать.

Список литературы

[Nur Arifin Akbar et al., 2021, Deep Learning of a Pre-trained Language Model's Joke Classifier Using GPT-2](#) - Файн-тьюнинг GPT-2 для генерации шуток, классификация получившихся шуток с помощью отфайн-тюнненного берта

[Yongyi Kui, 2021, Applying Pre-trained Model and Fine-tune to Conduct Humor Analysis on Spanish Tweets](#) - Использовали обученные модели (Albert-base-v2, XLNetbase-cased and Bert-base-multilingual-uncased) для классификации шутка/не шутка

[An Nguyen, 2021, Language Model Evaluation in Open-ended Text Generation](#) - были изучены разные метрики оценивания сгенерированного текста. Corpus-BLUE показывала наиболее близкие к человеческим оценки