DIALOGUE 2022
RUSSIAN TEXT DETOXIFICATION BASED ON PARALLEL CORPORA

РОМАН КАЗАКОВ, КСЕНИЯ ПЕТУХОВА, ВЕРОНИКА СМИЛГА

БКЛ182

Дано: датасет токсичных комментариев

Задача: нужно привести комментарии к нейтральному стилю

АКТУАЛЬНОСТЬ

- Социальная значимость: с ростом активности в социальных сетях распространяются такие явления, как травля и буллинг; с помощью систем детоксификации можно попытаться обезопасить общение незнакомых людей в чатах и комментариях
- Научная значимость: перенос стиля текста – комплексная задача NLP, задействующая разные уровни языка; качественное решение такой задачи стало возможным лишь с появлением сложных нейросетевых архитектур (например, CNN)
- Бизнес значимость: решение этой задачи может упростить модерацию социальных сетей, а также использоваться для контроля ответов диалоговых систем (чат-ботов и голосовых ассистентов)
- *Творческая значимость: кажется, что путей решения этой задачи может быть *дох*я -> очень много*

ОБЗОР ЛИТЕРАТУРЫ

- [Dale et al., 2021] – в работе представлены две модели для решения задачи детоксификации: 1) ParaGeDi: T5 для парафразы + GPT, чтобы выбирать самое вероятное, но при этом не токсичное слово; 2) CondBERT: детекция токсичных слов и подстановка вместо него слова, близкого по семантике
- [Dementieva el al., 2021] – первая работа, посвященная автоматической детоксификации русских текстов, в ней также представлены два подхода: 1) supervised подход на основе ruGPT-3 (команда "Перефразируй _ >>>"); 2) CondBERT, как в [Dale et al., 2021]

ДАННЫЕ

- **Параллельный датасет:** токсичное предложение на русском языке и 1-3 его нетоксичных аналога
- **Сообщения из соцсетей:** Одноклассники, Пикабу и Твиттер

Data	Кол-во токсичных предложений
Train	3539
Development	800
Test	1474

МЕТРИКИ ОЦЕНКИ

- **Style transfer accuracy (STA)**: бинарная метрика стиля, рассчитываемая с помощью классификатора токсичности на основе BERT, обученного на датасете русскоязычных токсичных комментариев (насколько удалось детоксифицировать)
- **Meaning preservation score (SIM)**: метрика косинусной близости, рассчитываемая с помощью эмбеддингов предложений LaBSE (насколько порождённое предложение сохранило семантику)
- **Fluency score (FL)**: метрика естественности, рассчитываемая на основе классификатора BERT, обученного на русскоязычных комментариях из социальных сетей и их автоматически сгенерированных аналогах (насколько текст похож на порождённый носителем языка)
- **Joint score (J)**: $STA \cdot SIM \cdot FL$
- Для финальной оценки private теста будет использоваться ручной аналог каждой из этих метрик: бинарная STA, бинарная SIM и трехклассовая FL

BASELINE

- **Delete-base**: удаление токсичных слов из предложений (на основе словаря). Наиболее примитивный вариант, не требует наличия параллельных корпусов

STA	0.53
SIM	0.87
FL	0.82
J	0.36

BASELINE

- **T5-base**: на трейне обучается модель Сбера ruT5-base. Наиболее стандартное решение при наличии параллельных данных, использует современную контекстуальную модель

STA	0.74
SIM	0.76
FL	0.9
J	0.45

ПЛАН

- Аугментация данных: можно расширить имеющийся корпус с помощью модели для перефразирования текста / с помощью подстановки синонимов на места токсичных слов
- Попробовать повторить эксперимент CondBERT из (Dale et al., 2021)
- Правиловая надстройка для слов и выражений типа *блин, бл*ть, еб*ать-копать* (их, например, можно просто удалять)
- Попробовать контекстуальные модели
- *Попробовать GPT
- Зафайтюнить модели под наши нужды (про все модели выше), надстройки над входными и выходными данными с использованием различных нейросетевых слоев

ЛИТЕРАТУРА

- Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova and Rada Mihalcea. “Deep Learning for Text Style Transfer: A Survey.” ArXiv abs/2011.00416 (2020)
- Daryna Dementieva, Daniil Moskovskiy, Varvara Logacheva, David Dale, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. “Methods for Detoxification of Texts for the Russian Language” Multimodal Technologies and Interaction 5 (2021): no. 9: 54. https://doi.org/10.3390/mti5090054
- David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov and Alexander Panchenko. “Text Detoxification using Large Pre-trained Neural Models.” EMNLP (2021)

Мы

- Рома: 1/3 R&D
- Ксюша: 1/3 R&D
- Ника: 1/3 R&D

- Мы пока изучаем статьи и придумываем, что делать, поэтому этот слайд откорректируем потом

Спасибо за внимание!