Гибкие диалоговые модели на основе метода байесовской оптимизации

Хуторной Дмитрий Павлович

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: dimon.khutornoy@mail.ru

Научный руководитель *— Ульянов Владимир Васильевич*

Работа посвящена **проблеме** отсутствия гибкости в моделях генерации текста. Множество алгоритмов остаются фиксированными после процесса обучения, и для их изменения приходится проделывать длительный и ресурсозатратный процесс переобучения.

Основной задачей является реализация гибкого диалогового ассистента, изменяющего свое поведение в соответствии с параметрами реплик, которые произносит собеседник. Данная задача является наглядным примером построения архитектуры гибкой модели.

В процессе работы были решены следующие **подзадачи**:

1. Проанализировать ряд алгоритмов классификации и генерации текстов и выбрать наиболее оптимальную архитектуру.

2. Собрать наборы данных и обучить модели.

3. Реализовать работу метода байесовской оптимизации совместно с имеющимися классификаторами и генератором.

4. Реализовать систему ранжирования сгенерированных ответов, а также механизм удаления наименее актуальных точек метода байесовской оптимизации.

Архитектура модели

1. 5 BERT-классификаторов характеристик реплик (токсичность, формальность, научность, политичность, эмоциональность) и BERT-классификатор удовлетворенности пользователя, позволяющий по тройке реплик ассистент-пользователь-ассистент оценить заинтересованность собеседника в диалоге.

2. GPT-модель для генерации нескольких вариантов ответа, которые затем ранжируются.

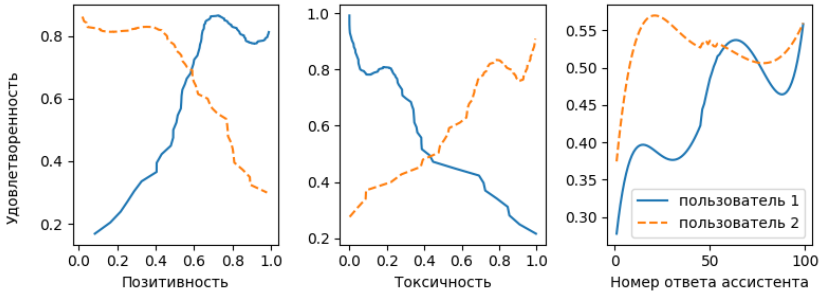
3. Байесовская оптимизация, функцией приспособленности которой является кусочное отображение, принимающее случайное

высокое значение, если реплика сказана пользователем, и результат классификатора удовлетворенности, если фраза исходит от ассистента.

4. Для большей гибкости из истории байесовского метода удаляются неактуальные точки.

В **результате** была получена модель, подстраивающаяся под внешние изменения и не требующая при этом много затрат на переобучение. Ниже приведены графики некоторых проекций функции приспособленности, характеризующие общение 2 пользователей с ассистентом (200 реплик). Видно, что поведение модели различается для каждого из них: первый предпочитает более грубое, а второй - позитивное. На третьем графике отображена тенденция удовлетворенности пользователей, которая имеет возрастающий тренд. Провалы на графике - моменты изменения поведения пользователя и соответствующего исследования новых точек функции приспособленности и удаления старых.

Иллюстрации



Сравнение двух пользователей модели

Литература

1. Ветров Д. П., Кропотов Д. А. Байесовские методы машинного обучения. 2007. URL: `https://bayesgroup.github.io/bmml/2016/BayesML-2007-textbook-1.pdf` (дата обращения: 01.04.2022)
2. Devlin J., Chang M., Lee K., Toutanova K., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 24.05.2019, arXiv:1810.04805.