Know What You Don't Know: Unanswerable Questions for SQuAD

(Pranav Rajpurkar, Robin Jia, Percy Liang)
Computer Science Department, Stanford University
11 June, 2018

Вопросно-ответные системы действительно часто дают верный ответ на поставленный вопрос, но только в том случае, когда нужная информация содержится в обрабатываемом тексте, однако, в обратной ситуации они порождают нерелевантные ответы, которых в действительности просто не существует. Имеющиеся датасеты также фокусируются на вопросах «с ответом» или используют автоматически сгенерированные вопросы «без ответа», которые легко распознать. Авторы статьи предлагают устранить обозначенную ими проблему, создав новый датасет SQuAD 2.0 (последняя версия the Stanford Question Answering Dataset). SQuAD 2.0 соединяет в себе уже существующую версию SQuAD 1.1 и более 50.000 вопросов «без ответа», созданных таким образом, чтобы они напоминали вопросы «с ответом». При работе с SQuAD 2.0 вопросно-ответные системы должны не только верно отвечать, когда искомая информация присутствует в тексте, но и воздерживаться от ответа, когда информация отсутствует.

Основными целями, которые авторы ставили перед собой при создании SQuAD 2.0, помимо базовых требований к датасету вроде большого размера, разнообразия и низкого шума, являлись также релевантность вопросов относительно темы рассматриваемого отрывка текста и наличие в нем правдоподобного ответа.. Прежде чем приступить к разработке SQuAD 2.0 , авторы статьи привели обзор существующих наборов данных, в котором выявили их сильные и слабые стороны, опираясь на установленные требования к их датасету.

Далее в статье описывался процесс формирования SQuAD 2.0. Для создания вопросов «без ответа» были наняты сотрудники, привлеченные через платформу краудсорсинга Daemo. В датасет были включены текстовые документы из SquAD 1.1 Для каждого абзаца в тексте было придумано пять вопросов, на которые невозможно было дать ответ. При этом вопросы соответствовали указанным требованиям (релевантность и правдоподобность) и напоминали вопросы «с ответом», а на работу с абзацем работнику отводилось семь минут. Из датасета были удалены вопросы, созданные теми краудсорсерами, которые придумали 25 и меньше вопросов к одному тексту, т.к. авторы считали, что это свидетельствует о недостаточном понимании задачи.

Для тестирования были использованы три модели: the BiDAF-No-Answer (BNA) и две версии Docu-mentQA No-Answer (DocQA). Для каждой модели был установлен определенный порог вероятности существования ответа на вопрос, дойдя до которого модель воздерживалась от ответа. При оценке на тестовой выборке использовался порог, максимизирующий F1 на выборке для обучения. В ходе эксперимента были поучены следующие результаты:

1) При тестировании моделей на SQuAD 1.1 точность выше, чем на SQuAD 2.0 (BNA: EM gap - 8.8, F1 gap – 15.2; DocQA: EM gap – 12.8, F1 gap – 18.7, DocQA + ELMo: EM gap – 15,2; F1 gap – 19,5)
2) Разница между точностью ответов, данными людьми, и тестируемой моделью с самыми высоким score намного больше в SQuAD 2.0 (SQuAD 1.1 test: EM – 3.7, F1 – 5.4; SQuAD 2.0 test: EM – 23.5, F1 – 23.2).

В заключение хочется отметить, что авторы статьи проделали огромную работу, в результате которой они представили сложный, разнообразный и крупномасштабный

набор данных, который стимулирует модель распознавать вопросы, на которые не могут быть даны ответы в рамках заданного контекста. Поставив перед собой такую задачу и успешно выполнив ее, они предоставили возможность создавать новые модели, которые будут «знать, то чего не знают» и поэтому понимать естественный язык на более глубоком уровне, что и является основной задачей автоматической обработки текста в настоящее время.