УДК 004.896

МЕТОДЫ ВОССТАНОВЛЕНИЯ QAT-ТРИПЛЕТОВ В УСЛОВИЯХ ОГРАНИЧЕННОСТИ ДАННЫХ

Александров Д. (федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – к.т.н., доцент ФЦТ, старший научный сотрудник НКЦР Бутаков Н.А.
(федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

В данной работе разработаны методы восстановления QAT-триплетов на основе ранжирования корпуса текстовых документов. Восстановленные триплеты в дальнейшем используются для обучения QAT-модели, входящей в состав вопросно-ответной системы.

Введение. В настоящее время поиск информации с одной стороны, становится более доступным, благодаря все увеличивающемуся числу источников информации и методов поиска среди них. Однако с другой стороны с увеличением количества доступной информации расширяется область поиска и нахождение ответов на специфичные вопросы становится затруднительным. Современные поисковые системы типа Google или Bing позволяют производить поиск по индексированным веб-документам, однако они не дают конкретные ответы на поставленные вопросы. Одним из активно развивающихся направлений вопросно-ответных систем области естественной обработки языка является задача нахождения ответа на вопрос в соответствующем тексте, под эту задачу обучаются QAT модели. Однако для обучения качественной QAT модели требуется большое количество данных в виде QAT триплетов. В реальной ситуации таких триплетов для конкретной области знаний крайне мало или зачастую не оказывается в наличии. Однако существуют источники специфичных знаний, такие как чаты, форумы, FAQ, которые содержат в себе пары вопрос-ответ с возможными ссылками на источники информации

Основная часть. QAT данные используются для обучения ранжированию документов из корпуса. Ранжирование нужно для восстановления связи между вопросом и документом. Ранжирование происходит в 2 этапа. На первом этапе с помощью методов без учителя отбираются топ-N релевантных документов. На втором этапе эти топ-N документов реранжируются более сложной моделью с учителем. Стоит отметить, что поиск документов ведется на всем корпусе: обучении и валидации. Можно выделить три метода поиска релевантных документов: по вопросу, по ответу, и одновременно по вопросу и ответу. Такое разделение позволит сравнить различные методы поиска и выбрать наиболее действенный.

После того, как к паре QA был подобран набор релевантных текстовых документов необходимо у них выделить ответы для полного восстановления QAT триплета. Может возникнуть иллюзия того, что нам уже доступен ответ в QA паре, однако ответ из данной пары не всегда может присутствовать в сопоставленном документе. Для нахождения ответа из текста путем выделения его части используются эвристики схожести между частью текста и ответа из QA пары. После применения эвристики, включающую в себя фильтрацию, на выходе получается восстановленный QAT триплет, который можно использовать для обучения QAT модели.

Исследования проводились на симулированных датасетах, предложен простой метод симуляции путем разделения изначального QAT-датасета на две части: QAT-части и QA-пар с набором документов. Процент QAT части определяет полноту такого датасета. С помощью определения уровня полноты датасета, при котором наблюдается существенное ухудшение качества QAT модели определяем полноту, при которой будут использоваться предложенные методы восстановления. Качество восстановления датасета, в том числе наличие шума в нем, сказывается на конечном качестве QAT модели.

BERT-QAT модель довольно устойчива к уменьшению числа обучающих данных (при сокращении обоих датасетов в 2 раза, качество модели по метрикам Exact Match и F1 упало примерно на 2%). Деградация модели наблюдается только при существенном сокращении обоих датасетов. Примем порог деградации модели для SQuAD в 5% полноты, так как при таком пороге наблюдается существенное ухудшение качества модели по метрикам Exact match (примерно 19%) и F1 (примерно 14%). Для датасета MS-MARCO примем порог полноты в 1%, наблюдается ухудшение метрики F1 примерно на 10% в абсолютной величие. В относительных величинах ухудшение качества на SQuAD при выбранном пороге – 16%, для MS-MARCO- 18%.

По лучшим и средним сеттингам восстановления проведены эксперименты по дообучению QAT моделей на восстановленных данных. Все предложенные методы weak-supervision показывают прирост в качестве QAT моделей на обоих датасетах. По метрикам качества восстановления датасета на результатах экспериментов на обоих датасетах лучшим показал себя метод с QA запросом. Однако на дообучении QAT моделей на восстановленных данных по методу с QA оказался хуже, чем метод только по вопросу: для SQuAD (+8.22 F1-score) и метода по ответу для MS-MARCO (+11.76 F1-score).

Выводы. Предложены методы восстановления QAT-триплетов на основе ранжирования текстовых документов. Проведены эксперименты по определению порога деградации модели, ранжированию и реранжированию документов, а также произведено дообучение QAT-модели на восстановленных триплетах. На обоих исследуемых вопросно-ответных датасетах (SQuAD и MS-MARCO) методы восстановления показали прирост в качестве QAT-модели.