УДК 004.85

СРАВНЕНИЕ ПОДХОДОВ К ПОЛУЧЕНИЮ КОНТЕКСТУАЛИЗИРОВАННЫХ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ СЛОВ ДЛЯ ЗАДАЧИ ВОПРОСНО-ОТВЕТНОГО ПОИСКА

Ефимов П.В. (Университет ИТМО, г. Санкт-Петербург)
Научный руководитель – к.т.н., доцент Муромцев Д.И.
(Университет ИТМО, г. Санкт-Петербург)

Работа посвящена сравнению эффективности различных подходов к получению контекстуализированных векторных представлений в многоязычных моделях на базе архитектуры Трансформер. Полученные векторные представления затем применяются для вопросно-ответного поиска, который является важной задачей обработки естественного языка. Тестирование моделей производится на русскоязычных частях датасетов XQuAD и TyDi QA.

Введение. Вопросно-ответный поиск (Question Answering) — область компьютерных наук, объединяющая обработку естественного языка и информационный поиск, целью которой является построение информационной системы, способной отвечать на вопросы пользователей на естественном языке. Изначально вопросно-ответный поиск выполнял задачи, связанные с удовлетворением информационных потребностей человека. Сейчас вопросно-ответный поиск может использоваться как формат для других задач: аннотирование текста, семантическая разметка, поиск скрытого смысла, визуальный вопросно-ответный поиск, использование как механизм переноса. Также вопросно-ответный поиск входит в наборы задач GLUE, SuperGLUE, RussianSuperGLUE, на которых оценивается способность современных моделей понимать естественный язык (Natural Language Understanding).

Долгое время большинство датасетов и инструментов в данной области были ориентированы на работу лишь с английским языком. В 2019 году возрос интерес к работе с многоязычными данными. Об этом говорит появление датасетов для тестирования моделей на различных языках (MLQA, XQuAD, TyDi QA).

Основная часть. Современные решения для задач обработки естественного языка основаны на языковой модели BERT (двунаправленное представление кодировщика из Трансформера). Использование модели BERT можно разделить на два этапа: предобучение (pre-training) и дообучение (fine-tuning). На этапе предобучения модель обучается на нескольких задачах на большом объёме неразмеченных текстов. Скрытые состояния BERT являются контекстуализированными векторными представлениями слов, которые на этапе дообучения можно использовать для решения конкретной задачи (например, вопросно-ответный поиск).

Многоязычная модель BERT (mBERT) — версия модели, обученная на 104 языках. Особенностью mBERT является то, что если дообучить её для какой-то задачи только на одном языке, то данная модель будет показывать приемлемые результаты и на других языках, на которых была обучена исходная языковая модель. Однако результаты могут различаться в зависимости от языка.

Качество межъязыкового переноса обучения можно повысить модифицировав контекстуализированные векторные представления с помощью следующих подходов:

1. изменение этапа предобучения — в отличие от BERT вместо пары предложений на одном языке на вход в модель подается два предложения на разных языках, и модель

должна научиться заполнять пропуски в предложении на одном языке используя слова предложения на втором языке (межъязыковая модель, XLM);
2. дообучение модели mBERT для задачи перевода контекстуализированных векторных представлений слов из одного языка в другой (Word-aligned BERT via fine-tuning);
3. преобразование векторных представлений слов из пространства одного языка в пространство другого перед применением их для вопросно-ответного поиска (Word-aligned BERT via rotation).

В предыдущих работах данные подходы хорошо зарекомендовали себя в других задачах обработки естественного языка, в частности в задаче логического вывода из текста и синтаксического анализа.

Эксперименты проводятся для межъязыкового переноса с английского языка на русский. Тестирование осуществляется на русскоязычных частях датасетов XQuAD и TyDi QA.

Выводы. Исследование подходов к получению многоязычных векторных представлений позволяет повысить качество межъязыкового переноса обучения между типологически разными языками. Также это позволяет получить новые сведения о знаниях, хранящихся в современных нейросетевых языковых моделях, на интерпретацию которых направлены многие текущие исследования. В дальнейшем планируется провести подобное сравнение для других языков.