УДК 004.89

ИЗВЛЕЧЕНИЕ АРГУМЕНТАЦИИ НА ОСНОВЕ ГЛУБОКИХ НЕЙРОННЫХ СЕТЕЙ ДЛЯ РУССКОЯЗЫЧНОГО КОРПУСА ТЕКСТОВ

Дудоладов С. (Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики)

Научный руководитель – д.т.н., доцент Котельников Е.В.

(Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики)

Целью доклада является показать существующие варианты решения проблемы извлечения аргументации и представить оптимальный в текущий момент набор параметров и модель для извлечения аргументации на русскоязычном корпусе.

Введение. Проблема автоматического извлечения аргументации является одной из ключевых задач обработки естественного языка в современном мире. Данная проблема была поставлена в 50-х гг. прошлого века в работе С. Тулмина, но каких-то значительных результатов исследователи смогли добиться только в последние годы с применением глубоких нейронных сетей и моделей на основе архитектуры Transformer. С другой стороны, работ, направленных на извлечение аргументации из текстов на русском языке, очень мало в связи с недостаточным количеством подготовленных корпусов размеченных текстов. Только в 2019 году И. Фищевой и Е. Котельниковым был получен первый русскоязычный корпус с аргументационной разметкой на основе перевода известного англоязычного корпуса ArgMicro. Поэтому задача исследования нейросетевых моделей извлечения аргументации для русского языка является актуальной и своевременной для изучения.

Основная часть. В ходе исследования были рассмотрены основные архитектуры, применяемые для извлечения аргументации, такие как рекуррентные сети, sequence-to-sequence, трансформеры, а также методы условных случайных полей и скрытые Марковские модели. Задача была поставлена как классификация с тремя классами аргументов: позитивными (поддерживающими основную точку зрения), негативными (возражающими основной точке зрения) и нейтральными (текстами, не содержащими позиции). Эксперименты проводились на переведенном корпусе ArgMicro. Все модели были оценены по основным метрикам классификации – точности, полноте и F1-мере. После этого на всех моделях был проведен подбор гиперпараметров, чтобы найти оптимальные параметры модели для русскоязычного корпуса.

Выводы. Результаты данной работы могут стать основанием для поиска оптимальной архитектуры для русского корпуса и на базе существующих моделей дают возможность разработать уникальную архитектуру для машинного извлечения аргументации из русскоязычных текстов. Данная архитектура может быть использована в различных прикладных системах – от рекомендации пользователю статей на основе его комментариев до систем принятия решений на бирже на основе анализа новостных статей и социальных сетей основных игроков рынка.

Дудоладов С.М. (автор) Подпись

Котельников Е.В. (научный руководитель) Подпись