УДК 004.822; 81’33

РАСПОЗНАВАНИЕ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ МЕЖДУ НАУЧНЫМИ ТЕРМИНАМИ

Тихобаева О.Ю. (Новосибирский национальный исследовательский государственный университет (НГУ)**)**

Научный руководитель – кандидат технических наук Бручес Е.П.

(Новосибирский национальный исследовательский государственный университет (НГУ))

Введение. На сегодняшний день уже существует огромное количество научных публикаций, представленных в электронном виде. И это количество продолжает расти. В связи с этим особую актуальность приобретают задачи, связанные с обработкой текстов научных статей. Становится необходимо автоматически извлекать из таких текстов различного рода семантическую информацию, в том числе семантические отношения между терминами. Их извлечение может быть полезно в отдельных специализированных областях, таких как поисковые и вопросно-ответные системы, а также при составлении онтологий.

Задача извлечения семантических отношений (Relation Extraction, RE) заключается в распознавании отношений, существующих между отдельными сущностями по контексту, в котором они находятся. В нашей работе в качестве сущностей мы рассматриваем научные термины, так как мы анализируем именно научные тексты.

В настоящий момент, данная задача остается сложной для любой предметной области, так как часто требует использования знаний вне текста (например, из баз знаний), а также из-за отсутствия большого количества размеченных данных на русском языке для решения задачи RE.

Целью данной работы является создание корпуса с разметкой семантических отношений для русского языка, а также реализация следующих алгоритмов: в классической постановке обучения с учителем и обучение без примеров (zero-shot learning).

Основная часть. Для создания корпуса взяли тексты из 10 предметных областей. В итоге нами было размечено 399 текстов, в которых было выделено 976 отношений 3 типов:

1) USAGE – ‘x используется для/в y’: 544;
2) ISA – ‘x является y’: 269;
3) PART_OF – ‘x является частью y’: 163.

Для тестирования алгоритмов мы использовали 20 текстов из изначального корпуса RuSERRC, а также по 20% текстов из новых размеченных данных для каждой предметной области.

Мы провели несколько экспериментов, попробовав различные подходы для извлечения семантических отношений.

Вначале, мы попробовали подход к извлечению отношений, который не требует обучения. Он заключается в следующем: если для каждой пары терминов составить предложения, отражающие все типы отношений (например, USAGE – “*мультимедийные технологии используются в учебном процессе*”) а затем получить оценку вероятности каждого из них с помощью модели GPT2 [1], то более вероятное предложение и будет отражать реальное отношение между терминами.

Далее мы попробовали использовать нейросетевую архитектуру, которая устроена следующим образом: берется вектор специального токена CLS (считается, что он представляет собой вектор всего текста, который пришёл на вход) и вектора двух наших терминов. Затем эти три вектора конкатенируются, и полученный вектор подается в классификатор [2].

Выводы. В ходе экспериментов с подходом, который не требует обучения, мы получили следующие метрики: точность/macro – 0.44, полнота/macro – 0.40, F1/macro – 0.39.

Подход, основанный на использовании нейросетевой архитектуры, показал следующие результаты: точность/macro – 0.75, полнота/macro – 0.71, F1/macro – 0.72.

В дальнейшем мы планируется провести эксперементы с другими методами извлечения отношений, которые не требуют использования большого количества обучающих данных.

Список использованных источников:

1. Radford A. et al. Language models are unsupervised multitask learners //OpenAI blog. – 2019. – Т. 1. – №. 8. – С. 9.
2. Wu S., He Y. Enriching pre-trained language model with entity information for relation classification //Proceedings of the 28th ACM international conference on information and knowledge management. – 2019. – С. 2361-2364.