ИИ «прочитал» книги о «Властелине колец» и научился анализировать литературу

Ученый из Института искусственного интеллекта AIRI и выпускница ВШЭ описали алгоритм автоматического анализа литературных произведений на основе ИИ через исследование системы персонажей книг Джона Рональда Руэла Толкина. Метод позволяет узнавать персонажей по их репликам и описаниям, а также определять характер их взаимоотношений. Технология применима для обучения диалоговых ботов и переводчиков, а также поможет сэкономить время при поиске информации в больших объемах текста.

Иллюстрации: нейросеть Stable Diffusion

Представьте, что вы читаете книгу. Вы без труда поймете, что какие-то слова в тексте – это имена, а какие-то – названия местности или организации, даже если впервые видите именно эти наименования. Для компьютерных систем распознавание имен людей, названий животных, организаций, топонимов – не самая простая задача.

Распознавание именованных объектов (Named Entity Recognition, NER) — это тип обработки естественного языка (Natural Language Processing, NLP), который помогает компьютерам идентифицировать и классифицировать объекты. NER похожа на цифрового библиотекаря, который настолько хорошо разбирается в литературе, что может прочитать текст и быстро выбрать из него самую важную информацию.

Используя технологии NER и теорию графов, старший научный сотрудник Института искусственного интеллекта AIRI Илья Макаров и выпускница ВШЭ Анастасия Ященко обучили алгоритм автоматического анализа литературных произведений на материале работ Джона Рональда Руэла Толкина и опубликованных после смерти писателя записях под редакцией его сына. В список вошли «История Средиземья», «Властелин колец» и «Хоббит». Система обучена распознавать именованные сущности, анализировать тональность текста и обнаруживать сообщества.

С помощью токенизации исследователи извлекли из текста 156482 предложения и в режиме «ручной настройки» получили список из 518 имен, 15 расовых лейблов и

биографических фактов. Например, «Бродда женился на Аэрин» или «Торин был убит Болгом». Этот список приняли за «золотой стандарт». Далее текст привели к нижнему регистру и убрали случаи перифраза и сокращений, заменив их универсальным именем (наиболее частым для персонажа). Позже список уточнили в автоматическом режиме, использовав более 4000 наиболее распространенных английских слов. Так, общее количество наименований для анализа составило 880 разных имен.

Для отображения связей между персонажами использовали пары имен, основанные на совместной встречаемости и обогащённые метаданными из индекса имён (семейные отношения, военно-исторические события и т.д.), а также анализ тональности предложений, в которых они появлялись.

Характер отношений между объектами описывали благодаря тональности контекста предложений, в которых упоминались оба героя. Например, если у двух персонажей по сюжету произведения были дружеские отношения, они склонны встречаться друг с другом в контексте вроде «они улыбались» или «обнимают друг друга», но никак не «они нападали друг на друга». Для финального деления использовали сумму тональности всех взаимодействий. Это позволило охарактеризовать отношения персонажей и описать сообщества в зависимости от того, насколько сильны взаимодействия между людьми. Выяснилось, что несмотря на обилие представленных в книге рас, система персонажей и их отношений в произведениях Толкина выражается не расовым разделением, а формируется на основе сюжетно близких групп, например, «Братство кольца».

Содержащийся в исследовании набор шагов можно использовать для извлечения именованных сущностей (названий, имен, мест и т.д.) и их взаимосвязей из других текстов. С помощью описанных методов можно выполнять ряд практических задач. Например, проводить анализ нормативной документации и суммаризацию юридических текстов, чтобы переписывать сложные документы простым и понятным обычному человеку языком.

«Практически все технологии можно отработать на базе не самых очевидных для бытовых представлений о серьезности науки примерах. В первую очередь, такой подход позволяет ученым получать удовольствие от исследований и обеспечивать более быстрый вход в изучение предмета для молодых специалистов. В прошлом году я уже руководил студенческой работой по автоматическому предсказанию сюжетных линий в 353 книгах о «Звездных войнах». Несмотря на взятые за основу научной статьи книги о галактических путешествиях, описанные в ней методы применимы для анализа исторических документов и создания аналитики огромных массивов текстовых данных по любой теме», – отметил **Илья Макаров**.

Вопросы: pr@airi.net