УДК 004.912

ИЗВЛЕЧЕНИЕ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ ДЛЯ СОЗДАНИЯ ОНТОЛОГИЙ ПРЕДМЕТНОЙ ОБЛАСТИ

Коробова П.И.
(Университет ИТМО)
Научный руководитель – к. т. н. Махныткина О.В.
(Университет ИТМО)

В данной работе рассматривается извлечение семантических отношений из неструктурированных или слабоструктурированных текстов лекций на русском языке для создания онтологий предметной области.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР № 620183 «Разработка виртуального диалогового помощника для поддержки проведения дистанционного экзамена на основе моделей-трансформаторов и понимания естественного и математического языка»

Извлечение терминов, определений, синонимов, гипонимов и других семантических отношений из неструктурированных текстов является важной задачей, которая используется для формирования баз знаний, тезаурусов и онтологий, для информационного поиска и других областей по обработке естественного языка.

В работе используются национальный корпус (а именно его синтаксически размеченный подкорпус «СинТагРус») и корпус научных текстов для русского языка RuSERRC. «СинТагРус» насчитывает около 77 тысяч предложений взятых из текстов различных жанров (художественная проза, научно-популярная литература, публицистика, биографии, статьи и новостные ленты). Он состоит из множества XML-файлов, разметка которых состоит из специальных тегов. RuSERRC представляет собой аннотации научных трудов в области информационных технологий, которые состоят из 1600 неразмеченных документов и 80 размеченных вручную документов, которые включают в себя шесть семантических отношений: причинно-следственная связь, отношение сравнения, таксономии, меронимии, синонимии и отношение использования. Наборы данных «СинТагРус» и RuSERRC применяются для обучения нейронной сети, а для файнтюнинга и тестирования используются предварительно размеченные в рамках проводимого исследования тексты лекций по дисциплине «Обработка естественного языка».

Для задачи классификации отношений используются векторные представления слов, полученные на основе языковой модели BERT, а именно RuBERT – это модель BERT, предварительно обученная на текстах русской Википедии. В качестве классификатора используются сверточные нейронные сети, которые подходят для извлечения семантически близких отношений, а именно для токенов, векторные представления которых находятся на небольшом расстоянии друг от друга (синонимы, гипонимы, гиперонимы), а вторая модель на основе рекуррентной нейронной сети LSTM подходит для токенов, векторные представления которых находятся на большом расстоянии друг от друга (термины, определения, меронимы, холонимы).

В данной работе был проведен анализ решений в области извлечения семантических отношений, осуществлен сбор и разметка набора данных, состоящего из лекций по дисциплине «Обработка естественного языка». Реализованы подходы к извлечению семантических отношений на основе машинного обучения.

Коробова П.И. (автор)

Махныткина О.В. (научный руководитель)