УДК 004.822

«Автоматизированное извлечение семантических отношений из текста с помощью Word2vec»

Автор: Ночевной Д. С., Университет ИТМО, Санкт-Петербург

Научный руководитель: Клименков С. В., Университет ИТМО, Санкт-Петербург

Формализовать знания в определённой предметной области и систематизировать их для быстрого доступа всегда являлось важной и актуальной задачей. Во многих научных областях такие системы знаний принято использовать для того, чтобы помочь пользователям легко и оперативно получать требуемую информацию. Одним из ключевых элементов систем автоматической обработки и хранения текста в упорядоченном виде являются онтологии или тезаурусы. Однако у всех подобных онтологий есть один недостаток – отсутствие специализированных терминов, специфичных для данной предметной области. Таким образом, появляется проблема дополнения существующей онтологии за счет добавления новых узлов и связей.

Целью работы является расширение исходной онтологии, в том числе восстановление отсутствующих связей для уже существующих узлов. Исходя из имеющихся исходных данных (начальная онтология и множество публичных ресурсов), а также поставленной цели, были выявлены задачи, требующие решения:

- автоматический поиск новых узлов, связанных семантическими отношениями (например, часть-целое или общее-частное);
- добавление новых узлов и связей в онтологию при обнаружении сущностей, признаки которых удовлетворяют заданным критериям.

Рассмотрено множество существующих решений в этой области, которые можно условно разделить на четыре категории:

- основанные на шаблонах (например, WHOLE “contain” PART для связи часть-целое);
- основанные на объектной модели документа (часто вложенные и родительские элементы DOM имеют семантическую связь);
- основанные на форматировании текстовой информации (в частности, таблицы);
- основанные на машинном обучении.

В нашем предыдущем исследовании был описан и реализован метод, основанный на анализе структуры текстовой информации. С помощью этого подхода были получены достаточно точные результаты для форматирования с помощью таблиц, списков, а также специальных шрифтов.

В данной работе рассматривается использование инструмента Word2vec, так как он позволяет преобразовать слова в векторное (числовое) представление и проанализировать их смысл и контекст, в котором они используются. Таким образом, в данном случае можно говорить о получении наиболее вероятного смысла для каждого слова в тексте. К задачам, решаемым Word2vec, можно отнести кластеризацию, поиск семантически близких слов. Но в данной работе ключевым источником для извлечения информации служит разность векторов слов, которая является не чем иным, как отражением связи между ними.

В процессе настоящего исследования был разработан специальный текстовый корпус, а также программное обеспечение, которое позволило как восстановить связи между уже существующими узлами онтологии, так и создать новые сущности.

В настоящее время использование машинного обучения для получения семантических отношений только зарождается. Было найдено лишь небольшое количество исследований, посвященных этой теме. Полученные результаты показывают необходимость продолжать исследования. В дальнейшем планируется создание нейронной сети, которая сможет принимать в качестве входных данных не только обычный текст, но и частично

структурированную текстовую информацию, а также поиск способов повышения точности использованных алгоритмов.

Автор _____/_____(Фамилия И.О.)

(подпись)

Научный руководитель _____/_____(Фамилия И.О.)

(подпись)

«__27__» __февраля__2019 года