УДК: 004.822

АВТОМАТИЧЕСКОЕ ОБНОВЛЕНИЕ СЕМАНТИЧЕСКОЙ СЕТИ ИЗ ОТКРЫТЫХ ИСТОЧНИКОВ

Зубань Д.А. (Национальный исследовательский университет ИТМО), **Покид А.В.** (Национальный исследовательский университет ИТМО), **Яркеев А.С.** (Национальный исследовательский университет ИТМО)
Научный руководитель - ассистент Письмак А.Е. (Национальный исследовательский университет ИТМО)

В данной работе описан подход инкрементального обновления семантической сети данными из открытого словаря ВикиСловарь. В результате работы удалось реализовать механизм периодического обновления семантической сети и повысить качество работы алгоритмов обработки естественного языка на текстах содержащих предметно специфичную и новую лексику.

Введение.

Извлечение информации из слабоструктурированных источников является одним из приоритетных направлений компьютерной лингвистики. Огромный рост данных данных хранящихся в слабо структурированном виде требует специализированных инструментов для автоматической извлечения информации из текста с целью её дальнейший обработки. Область применения обработки естественного языка огромна (машинный перевод, построение экспертных систем, выявление информационных атак, автоматическое реферирование).

Для решений задач обработки естественного языка используются различные языковые модели разных типов (n-граммы, нейронные сети, “bag-of-concepts”, семантические сети и онтологии). Семантические сети представляют собой способ представления знаний в виде размеченного ориентированного графа. С помощью этого графа описывается иерархия понятий и определяются отношения между понятиями (меронимия, холонимия и др.).

Основная часть.

В данной работе рассматривается процесс автоматического обновления семантической сети. Построение семантической сети требует большого объема работ и привлечения квалифицированных лингвистов и экспертов различных предметных областей. Существует также подходы автоматического построения семантических сетей. Недостатком такого подхода является требование контроля качества к полученным результатам. В ходе предыдущих работ авторами была создана семантическая сеть для русского языка основанная на открытых источниках. Одним из взятых за основу источников является лингвистическая онтология “РуТез”. Тезаурус Рутез был вручную построен коллективом профессиональных лингвистов. Преимуществом данного источника является хорошо проработанная ядро семантических понятий языка. Недостатком данного ресурса является слабый охват различных предметных областей. Для решения этой проблемы для построения семантической сети использовался второй источник.

Вторым источником для построения сети является Викисловарь. Данный ресурс представляет из себя свободно пополняемый многоязычный словарь. Викисловарь содержит огромное количество понятий из разных предметных областей и постоянно пополняется новыми данными большим количеством энтузиастов и развитым сообществом. Однако данный словарь не является семантической сетью и семантические отношения присутствующие в словаре связывают слова, а не понятия. Однако в рамках предыдущей работы авторы разработали алгоритмы по извлечению необходимых данных из викисловаря и преобразованию их к семантической сети.

В ходе данной работы был создан алгоритм автоматического обновления семантической сети. В работе был изменен алгоритм импорта данных из Викисловаря, реализованный в исходной версии системы управления семантической сетью. Необходимость изменения работы алгоритма импорта данных из Викисловаря обусловлена постоянным дополнением и внесением исправлений в словарь сообществом. В результате работы был получен модуль периодического импорта обновлений из словаря. Обновления семантической сети происходят инкрементально и не требуют процесса повторного построения сети с нуля.

Выводы.

В ходе работы удалось реализовать алгоритм периодического обновления семантической сети. Данное изменение позволит использовать при работе алгоритмов обработки текста на естественном языке информацию о новых понятиях и связях добавленных в языковую модель. Актуализация языковой модели является неотъемлемой задачей в связи с постоянной эволюцией естественного языка и возникновением новых понятий в языке.

Яркеев А.С. (автор)

Письмак А.Е. (научный руководитель)