УДК 004.822

ФОРМИРОВАНИЕ СЕМАНТИЧЕСКОЙ СЕТИ АДРЕСНОЙ ИНФОРМАЦИИ ИЗ ОТКРЫТЫХ ИСТОЧНИКОВ

Рогаленко Н.А. (Университет ИТМО)
Научный руководитель – аспирант Яркеев А.С.
(Университет ИТМО)

В докладе рассматриваются проблемы обработки неструктурированных данных на естественном языке и использование системы семантического анализа текста для их решения. В качестве предметной области исследования было выбрано адресное пространство городов и населенных пунктов. Основная часть доклада посвящена процессу автоматического создания семантической сети адресной информации, включая обзор возможных источников данных, описание архитектуры программной системы и алгоритмов ее работы.

Введение.

Неструктурированность и неоднозначность естественного языка являются источником множества проблем при обработке данных, поскольку допускают неточности в трактовке. Для формализации таких данных появляется необходимость искать в тексте на естественном языке объекты определенного класса. При этом важно учитывать, что в тексте слова, описывающие объект, могут встретиться в разных формах, в разных падежах, в разном порядке. С усложнением структуры объекта увеличивается и сложность его описания, из-за чего найти в тексте объект со сложной структурой с помощью регулярных выражений становится затруднительно.

Альтернативным решением может послужить система семантического анализа текста, которая решает задачу поиска в тексте на естественном языке «именованных сущностей», т.е. конкретных объектов заданного класса. По словам в тексте определяется, что это за объект, благодаря чему появляется возможность получить дополнительную информацию о нем. В ходе исследования в качестве предметной области для представления в виде семантической сети выбрано адресное пространство, т.е. классом для распознавания является адрес, а объектом - конкретный адрес. Подобный выбор обусловлен тем, что распознавание объектов этого класса может быть полезно во множестве систем – в логистике, при валидации веб-форм, при работе с различными документами и др.

Основная часть.

Для использования системы семантического анализа необходимо наполнить семантическую сеть данными, поэтому первоочередной задачей был поиск источника адресной информации. Среди критериев, по которым выбирался источник, можно выделить открытость, структурированность данных, достоверность и обновляемость, что особенно важно, поскольку адресные объекты часто переименовываются и теряют свою актуальность. В качестве источника данных были рассмотрены источники трех категорий – картографические (openstreetmap, yandex map, google map), базы знаний (Wikidata, DBpedia), официальные государственные реестры РФ (КЛАДР, ФИАС). В итоге была выбрана федеральная информационная адресная служба (ФИАС), как наиболее достоверный и полный источник. Выгрузки данных производятся в формате XML, содержат информацию об адресных объектах, домах, квартирах, что позволяет в полной мере описать адрес. Объекты представлены в виде тегов с набором атрибутов для идентификации объекта, определения актуальности и прочих характеристик. Информация обновляется два раза в неделю.

Для реализации задачи наполнения семантической сети адресной информацией был разработан комплекс программ. В первую очередь - подсистема обработки файлов ФИАС, определяющая, какие адресные объекты являются актуальными, и какие словоформы можно отнести к объекту. Полученные данные в промежуточном виде передаются подсистеме генерации скрипта, содержащего предложения на специальном декларативном языке SemQL, предназначенном для работы с семантическими сетями. Полученный скрипт обрабатывается модулем импорта данных в семантическую сеть, наполняя ее словоформами, смысловыми понятиями и их экземплярами. Кроме того, поскольку адресная информация постоянно обновляется, были добавлены модуль автоматической выкачки файлов ФИАС и подсистема обновления, анализирующая текущее состояние сети и определяющая, какие изменения необходимо ввести при обновлении адресной базы, что позволило генерировать SemQL-предложения для изменения уже импортированных адресов и избавило от необходимости пересоздавать всю сеть при каждом обновлении.

Выводы.

В конечном итоге была реализована система, осуществляющая обработку файлов ФИАС и генерирующая на их основе скрипт на языке SemQL, обрабатываемый специальным модулем импорта, с помощью которого семантическая сеть автоматически наполняется адресными данными. При этом адресная база постоянно поддерживается в актуальном состоянии, а благодаря сгенерированным семантическим отношениям и словоформам система позволяет достичь высокой гибкости при поиске адресных объектов в тексте на естественном языке. Как итог работы системы была получена семантическая сеть адресной информации, содержащая более сотни смысловых значений и более миллиона экземпляров и словоформ.

Рогаленко Н.А. (автор) Подпись

Яркеев А.С. (научный руководитель) Подпись