Разработка и реализация алгоритма извлечения онтологии из научного текста произвольной структуры

Пастушков Илья Сергеевич
Новосибирский государственный университет (Новосибирск), Россия
e-mail: pas2shkov.ilya@gmail.com

Проблема поиска информации — одна из вечных проблем человеческого сообщества. На протяжении своего многотысячелетнего развития его представители неустанно находятся в поиске того, где находится что-либо: пища, жилище, пастбища, дороги, сокровища и т. п. Обобщая задачи поиска, можно сказать, что человечество постоянно находится в поиске знаний.

В извлечении фактов проблемой является то, что даже если термин употребляется в документе, не факт, что из этого можно извлечь какую-либо полезную информацию.

Для извлечения фактов из текста необходимо извлечь онтологию, чтобы на её основе реализовать алгоритм фактографического поиска. В извлечении онтологии и состоит задача.

На сегодняшний день существует большое количество подходов для обработки естественного языка, но в большинстве своём, они не подходят для данной задачи. Проблема в том, что английский, для которого, в основном, такие системы создаются, как и большая часть романо-германских языков, не обладает настолько сложной морфологией, как русский.

Система извлечения морфологий должна основываться на обучаемом алгоритме без учителя, поскольку даже тексты авторов одного направления могут обучить алгоритм противоположным вещам.

В данной работе, на основе морфологического анализа текста с помощью метода опорных векторов и словаря корпуса русского языка, а также последующего применения метода случайных блужданий был получен список семантически связанных слов, опираясь на который можно построить онтологию для конкретного документа, что в свою очередь является основой для реализации алгоритма фактографического поиска.

В качестве результата данной работы:

- Разработан и реализован морфологический анализатор для русского языка ;
- Разработан и реализован алгоритм извлечения слов с заданной семантической связностью;
- Проведена исследовательская работа по изучению различных алгоритмов фактографического поиска.

Разработка производилась на языке Python.

В перспективе:

- На основе полученных результатов реализовать алгоритм фактографического поиска
- Сделать поиск масштабируемым, что позволить проводить поиск более, чем по одному документу

- Оптимизировать исходный код для большей производительности

Литература

1. L.P. Coelho, W.Richert, Building Machine Learning Systems with Python
2. К.Д. Маннинг, П.Рагхаван, Х.Шютце, Введение в информационный поиск, Вильямс, ISBN 978-5-8459-1623-5, 978-0-5218-6571-5; 2011 г.