Д. А. Свядыш

АНАЛИЗ СУЩЕСТВУЮЩИХ СРЕДСТВ ИЗВЛЕЧЕНИЯ СТРУКТУРИРОВАННЫХ ДАННЫХ ИЗ ТЕКСТА НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Рассматриваются средства извлечения структурированных данных из текста на естественном языке. Проводится сравнительный анализ таких средств.

Введение

Извлечение фактов из текста — типичная задача при работе с естественным языком. В современных информационных технологиях роль такой процедуры, как извлечение информации, всё больше возрастает — из-за стремительного увеличения количества неструктурированной информации. В данной статье рассмотрим основные принципы работы двух средств извлечения структурированных данных из текста, а также сравним их.

I. Томита-парсер

Томита-парсер позволяет по написанным пользователем шаблонам (КС-грамматикам) выделять из текста разбитые на поля цепочки слов или факты [1].

Основные компоненты парсера: газеттир, набор грамматик и множество описаний типов фактов, которые порождаются этими грамматиками в результате процедуры интерпретации.

Газеттир — словарь ключевых слов, которые используются в процессе анализа КС-грамматиками. Каждая статья этого словаря задает множество слов и словосочетаний, объединенных общим свойством.

Грамматика — множество правил на языке КС-грамматик, описывающих синтаксическую структуру выделяемых цепочек.

Факты — таблицы с колонками, которые называются полями фактов. Факты заполняются во время анализа парсером предложения.

Помимо того, что Томита-парсер успешно справляется с поставленными задачами по извлечению структурированных данных из текста, данная технология имеет широкие возможности кастомизации.

II. Cloud Natural Language

Google Cloud Natural Language API выявляет структуру и значение текстов, используя мощнейшие модели на базе технологий машинного обучения, которые упрощают работу с REST API [2]. Таким образом, разработчики получают возможность использовать данные о людях, местах, событиях и прочих реалиях, которые когда-либо упоминались в текстах новостей, статей и блогов. Также данная технология позволяет классифицировать статьи (см.рис.1.)

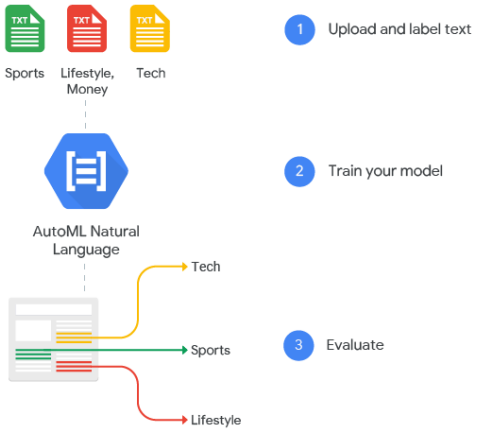

Рис. 1 – Схема работы Cloud Natural Language

Таким образом, Cloud Natural Language предоставляет удобный интерфейс для анализа текстов, а также имеет большую базу данных о реальном мире.

III. Выводы

Рассмотренные средства извлечения структурированных данных из текста позволяют работать с статьями, написанными на естественном языке, однако имеют некоторые отличия: Томита-парсер имеет широкие возможности по кастомизации, в то время как Cloud Natural Language не обладает такой гибкостью, но в то же время использует большие базы данных при обработке текста.

1. Томита-парсер – Технологии Яндекса [Электронный ресурс] / Яндекс. – Москва, 2019. – Режим доступа: https://tech.yandex.ru/tomita/. – Дата доступа: 25.02.2019.
2. Cloud Natural Language [Electronic resource] / Google. – Mountain View, California, 2019. – Mode of access: https://cloud.google.com/natural-language/. – Date of access: 26.02.2019.

Свядыш Дмитрий Алексеевич, магистрант кафедры интеллектуальных информационных технологий БГУИР, dmitry.svyadysh@gmail.com.

Научный руководитель: Сердюков Роман Евгеньевич, кандидат технических наук, доцент кафедры интеллектуальных информационных технологий БГУИР, rserdyukov@gmail.com.