Распознавание именованных сущностей с использованием синтактико-семантических признаков и нейросетей

МФТИ

Юсупов Идрис (i.yusupov@phystech.edu)

1. Цель работы

Разработать новостной агрегатор, который соответствует следующим требованиям:

- автоматическая классификация новости к городу,
- всё должно происходить без участия человека, но должна быть возможность вмешаться,
- пока для **5 городов**.

2. Общая схема

3. Парсер

Извлекает из текста:

- длиннейшие последовательности слов с большой буквы,
- однокоренные с городом слова(см. правила в словаре),
- слова в кавычках.

4. Вектор признаков

- Мешок слов,
- доменное имя источника новости,
- фильтруется с помощью словаря.

5. Словарь

В словаре хранятся токены и правила.

Токен:

- нормализован(стемминг),
- относится к городам.

Правило:

- на основе регулярных выражений,
- может отображаться в токен,
- относится к городам.

Начальные данные извлекаются из OpenStreetMap(названия улиц, достопримечательности и т.п).

6. Классификаторы

В качестве классификатора используется **Multinomial Naive Bayes with ROSE-smoothing**[1].

Классификатор по городам

- "один против всех",
- количество классификаторов получается равным количеству городов.

Классификатор выбросов

- Необходимо определять новость, которая не относится к представленным городам.
- Для этого используем классификатор по 2м классам - "относится к городу" и "НЕ относится к городу".

7. Извлечение отношений

1. В классифицированном документе находятся пары вида(ГОРОД, СЛОВО_ИЗ_СЛОВАРЯ_ГОРОДА).
2. По найденным вхождениям создается *вектор признаков*(левый и правый контексты, позиция слова и др.).
3. На основе сформированных вектров признаков создаются паттерны.
4. По паттернам находим новые слова.

Похоже на DIPRE[2].

8. Результаты

9. Заключение

Разработана расширяемая система:

- с использованием машинного обучения,
- с начальным заполнением словаря геоданными из OpenStreetMap,
- оценка точности: > 85%,
- сайт: rbcitynews.ru.

В будущем планируется:

- добавление новых городов,
- извлечение парсером имён собственных,
- кластеризация токенов в словаре.

Благодарности

- Научному руководителю Макееву Г.А.
- Компании "Техинформ".

Ссылки

1. Alexander Y. Liu and Cheryl E. Martin. 2011. Smoothing multinomial naïve bayes in the presence of imbalance
2. Sergey Brin. 1998. Extracting Patterns and Relations from the World Wide Web.
3. http://www.openstreetmap.org/
4. http://mit.spbau.ru/files/datamining1009.pdf