Летняя международная XXXI молодежная Школа-конференция по параллельному программированию

Извлечение информации из текста на естественном языке

Темникова Екатерина НГУ ФИТ 1 курс

Пчелкина Алиса НГТУ ФПМИ 1 курс

Руководители: компания Dasha.AI

13 июля 2018

План доклада

- Кратко о NLP (определение, задачи)
- Задача проекта
- Идея реализации
- Кратко о Томита-парсере.
- Пример правила грамматики
- Пример входных и выходных файлов парсера

Обработка естественного языка

- Википедия

Обработка естественного языка(*Natural Language Processing,NLP*)

— общее направление искусственного интеллекта и математической лингвистики. Оно изучает проблемы компьютерного анализа (понимания языка) и синтеза (генерация грамотного текста) естественных языков.

Задачи и ограничения

1. Распознавание речи
2. Анализ текста
   - А) Извлечение информации
   - Б) Информационный поиск
   - В) Анализ высказываний
   - Г) Анализ тональности текста
   - Д) Вопросно-ответные системы
3. Генерирование текста
4. Синтез речи

Задачи и ограничения

Понимание естественного языка иногда считают AI-полной задачей, потому как распознавание живого языка требует огромных знаний системы об окружающем мире и возможности с ним взаимодействовать. Само определение смысла слова «понимать» — одна из главных задач искусственного интеллекта.

AI-полная задача – проблема, решение которой предполагает решение главной проблемы искусственного интеллекта — сделать компьютеры такими же умными, как люди

Задачи проекта

- Извлечь и классифицировать информацию о дате и времени или о промежутках времени из неструктурированного текста.

Вчера был дождь	< вчера >
Позвоните мне через два дня	<через><два><дня>
Увидимся завтра!	<завтра>
Мне удобно будет после обеда	<после><обеда>

Задачи проекта

- Извлечь и классифицировать информацию об адресах из неструктурированного текста.

Я был в Москве < Москва >

614000 г Пермь Шоссе Космонавтов

<61400><г><Пермь><шоссе><Космонавтов>

Ленина 5/2 7 <Ленина><5/2><7>

Ул Коммунистическая 5 корпус 4 строение 1 офис 17

<ул> <Коммунистическая> <5><корпус 4 строение 1><офис><17>

Идея реализации

- Использовать GLR-парсер для быстрого и эффективного распознавания текстов, написанных на естественном языке. Обычный LR-парсер не способен разрешать недетерминированность и неоднозначность естественных языков, тогда как GLR-алгоритм может.

Томита-парсер

Томита-парсер создан для извлечения структурированных данных из текста на естественном языке. Вычленение фактов происходит при помощи контекстно-свободных грамматик и словарей ключевых слов. Парсер позволяет писать свои грамматики и добавлять словари для нужного языка.
Исходный код проекта открыт и выложен на GitHub.

Руководство разработчика >

Видеокурс >

Репозиторий >

Открытый код Томита-парсера и его использование вне Яндекса

Принцип работы

- Томита-парсер позволяет по написанным пользователем шаблонам (КС-грамматикам) выделять из текста разбитые на поля цепочки слов или факты. Например, можно написать шаблоны для выделения адресов. Здесь фактом является адрес, а его полями — «название города», «название улицы», «номер дома» и т.д.
- Парсер включает в себя три стандартных лингвистических процессора: токенизатор (разбиение на слова), сегментатор (разбиение на предложения) и морфологический анализатор (mystem).
- Основные компоненты парсера: газеттир, набор КС-грамматик и множество описаний типов фактов, которые порождаются этими грамматиками в результате процедуры интерпретации.

Правила грамматики

S —> S1 ... Sn { Q };

Нетерминал --> терминал или нетерминал;

Порядок перечисления правил.

От порядка перечисления правил работа программа не зависит.

Если во входном тексте находится цепочка, которая соответствует правой части, то правило «срабатывает» и грамматика использует эту цепочку как значение символа LeftPart

Пример

```
MoscowWord -> "москва"<h-reg1>;
MoscowGroup -> Adj<gnc-agr[1]>* MoscowWord<rt,gnc-agr[1]>;
```

Примеры входных и выходных данных

198332 г Санкт-Петербург проспект Маршала Жукова д 37 кв 317 462404 Оренбургская область г Орск Суворова д 26

195273 г Санкт-Петербург Верности д 48 кв 63

420044 Казань Короленко д 67 кв 32

350061 г Краснодар Трудовой Славы ул д 19/2 кв 57

614000 г Пермь Шоссе Космонавтов д 213 кв 8

420100 Казань Тыныч ул д 3 кв 8

площадь ленина 1 /3 корпус 7

Кутузова д 24 офис 7

просп Будёнова д 62 стр 2

Примеры входных и выходных данных

Address

Index	RegionD	RegionName	CityD	CityName	StreetD	StreetName	HouseD	HouseNumber	ComplexHouseNumber	FlatD	FlatNumber
198332			г	Санкт-Петербург	проспект	Маршала Жукова	д	37			
										кв	317
462404	область	Оренбургская	г	Орск		Суворова	д	26			
195273			г	Санкт-Петербург		Верности	д	48		кв	63
420044				Казань		Короленко	д	67		кв	32
350061			г	Краснодар	ул	Трудовой Славы	д	19/2		кв	57
614000			г	Пермь		Шоссе Космонавтов	д	213		кв	8
420100				Казань	ул	Тыныч	д	3		кв	8
					площадь	ленина		1 / 3	корпус 7		
						Кутузова	д	24		офис	7
					просп	Будёнова	д	62	стр 2		

Заключение

В результате проделанной работы все поставленные задачи были достигнуты