Московский государственный университет имени М.В. Ломоносова

Факультет вычислительной математики и кибернетики

Домашнее задание

Извлечение фактов

Выполнили:

Смирнова Александра Сергеевна

Киямова Александра

Москва, 2016

Оглавление

1. Описание алгоритма

1.1. Постановка задачи

Исходные данные: текст с выделенными именными сущностями типа "Org"(организация), "Person"(персона), "Job"(работа/вид деятельности).

Задача: выделить из текста факты типа "Occupation"(занятость человека на работе). Данный факт имеет одно обязательное поле: "Person", и два необязательных: "Job" и "Org".

1.2. Алгоритм

Для решения поставленной задачи было решено использовать инженерный подход, а именно: извлекать факты при помощи шаблонов. На основе обучающей выборки текстов были вручную составлены шаблоны, которые оперируют единицами Org, Person и Job. Для извлечения данных единиц так же использовались шаблоны, но уже сгенерированные автоматически на основе выделенных в тексте именованных сущностей.

В оригинальной постановке задачи поля факта "Job" и "Org" являются необязательными, но составленные правила ориентируются только на факты с одним необязательным полем "Org".

Набор шаблонов для извлечения фактов описан ниже (Word - любое слово).

- F -> Job Person
- F -> Job Org Person
- F -> Job Org Word Person
- F -> Job Org Word Word Person
- F -> Job Org Word Word Word Person
- F -> Person Job Org
- F -> Person Job Word Org
- F -> Person Job Word Word Org
- F -> Job Word Org Person
- F -> Job Word Word Org Person
- F -> Job Word Word Word Org Person

2. Описание программы

В качестве инструмента для описания шаблонов использовался Томита-парсер. Для работы с данным инструментом на языке парсера были описаны все необходимые файлы конфигурации. Факты извлекаются на основе четырех файлов с грамматиками. Первый файл содержит вручную составленные правила для извлечения непосредственно самого факта, остальные три файла содержат грамматики для извлечения необходимых именованных сущностей. Именно эти три файла автоматически генерируются для каждого текста корпуса. На Рисунке 1 показан пример составленной грамматики для извлечения именованной сущности типа "Org".

```
#encoding "utf-8"
#GRAMMAR_ROOT Org

Org ->  "минприроды" "россия";
Org ->  "госкомпания" "по" "геологоразведка" "росгеология";
Org ->  "минэнерго";
Org ->  "государственный" "комиссия";
Org ->  "государственный" "компания" "по" "геологоразведка" "росгеология";
Org ->  "россия";
Org ->  "советский" "союз";
```

Рисунок 1. Пример грамматики

Для автоматической генерации грамматик и запуска Томиты-парсера была написана программа на языке Python. При написании данной программы для работы с корпусом текстов использовался парсер, предоставленный Артемом Казаковым.

3. Методика тестирования

Для тестирования использовался инструмент, прилагающийся к корпусу текстов. Данный инструмент считает полноту, точность и F-меру извлеченных фактов.

4. Примеры работы программы

Для каждого текста корпуса программа составляет файл особого формата, который содержит извлеченные факты. Каждый факт имеет вид:

Occupation
Who: <имя работника>
Where:<название организации>
Job:<должность/вид деятельности>

Строка Where может отсутствовать.

Например, для текста, содержащего следующие фрагменты: "... так начинается доклад американского нейробиолога Гэри Уилсона на конференции TED ... президент США Джон Кулидж отпустил при посещении ..." были извлечены следующие факты:

Occupation
who:гэри уилсона
job:американского нейробиолога

Occupation
who:кулидж
where:сша
job:президент

5. Описание результатов и их анализ

Программа была протестирована дважды. В первый раз для извлечения фактов использовался только один шаблон: F -> Job Org Person. Во второй раз использовался полный набор шаблонов, описанный в 1 разделе. В таблице 1 описаны полученные результаты.

	Точность	Полнота	F-мера
1 шаблон	85%	30%	44%
Полный набор шаблонов	63%	43%	51%

Таблица 1. Результаты работы программы

По результатам первого тестирования можно увидеть, что большая доля фактов покрывается одним простым шаблоном, при этом точность держится на довольно высоком уровне.

При добавлении других шаблонов полнота возрастает на 13%, а точность при этом падает на 22%. При ручном просмотре извлеченных фактов было замечено, что при втором тестировании прирост полноты идет в основном за счет шаблона: F -> Job Person. Однако в то же время, именно факты, извлеченные при помощи данного шаблона, так негативно повлияли на точность.

Как следствие, один из способов улучшить результаты программы - это модифицировать правило F -> Job Person. Возможно, если конкретизировать некоторые слова вокруг данной конструкции, то точность не будет так сильно падать.

Второй способ улучшения - это добавить новые шаблоны, которые бы учитывали ситуации, где поля факта находятся на большом расстоянии друг от друга, возможно даже в разных предложениях. Но как и в случае с шаблоном F -> Job Person, скорее всего придется уделить внимание более точному описанию окружения полей факта, чтобы не получить резкое уменьшение точности.