Обработка естественного языка, лекция 2: вспомогательные задачи и формальные грамматики

Пожидаев М. С.

17 ноября 2021 г.

Токенизация

Токенизация — преобразование входной строки для выделения групп символов, формирующих токены.

1. Точные требования к токену определяются задачей (включение пробелов, дробные части слов и пр.).
2. Иногда проводится аннотирование и распознавание понятий (цен, дат, инициалов и пр.).
3. В большинстве случаев является первым этапом обработки текста.

Сегментация

Сегментация — разделение текста (или последовательности токенов) на синтаксически обособленные группы (между сегментами нет синтаксических связей, только семантические).

1. Появление точки, вопросительного знака или восклицательного знака с последующим словом, начинающимся с заглавной буквы, — необходимое условие начала нового сегмента (предложения), но не достаточное!
2. Сегментация и токенизация в общем случае являются многофакторным задачей, требующей учёта множества различных аспектов (которые даже человеку не всегда легко разобрать), поэтому известно множество различных подходов к их решению, включая использование искусственных нейросетей.

Распознающие формальные грамматики

1. Задаются алфавит, множество терминальных символов и нетерминальных символов.
2. Задаётся множество правил, у которых в левой части находится один нетерминал, а в правой последовательность из терминальных и нетерминальных символов.
3. Задаётся корневой нетерминальный символ.

Требуется определить дерево правил, терминальные символы которых совпадают с распознаваемой последовательностью.

Томита-парсер

Томита-парсер (*Tomita-parser*) — открытая утилита компании Яндекс для обработки естественного языка.

1. Производит извлечение фактов из текста.
2. Текст анализируется посредством распознающих формальных грамматик.
3. Использование затруднено большой трудоёмкостью обработки текста при недетерминированных правилах и сложностью расстановки приоритетов.

Основные понятия Томита-парсера

1. Газеттир — множество контекстносвободных правил.
2. Факт — запись с атрибутами, описывающая распознанное понятие.
3. Грамматика — способ связи правил и фактов.

Пример грамматики

```
Ent -> 'свод' 'череп' ;
Ent -> Adj<gnc-agr[1]>+ Noun<gnc-agr[1], rt> ;
Rel -> Ent interp (Equality.Name1) "являться"\
     Ent interp (Equality.Name2) ;
Rel -> Ent interp (Equality.Name1) "образовывать"\
     Ent interp (Equality.Name2) ;
```

Лемматизация и стемминг

Лемматизация — получение словарной формы словоформы.
В расширенном смысле с получением морфологических атрибутов.
Исходный текст: *Я ходил в магазин за булочкой.*
Лемматизированный текст: *Я ходить в магазин за булочка.*

Стемминг — выделение основы слова без окончания.
Результат недетерминирован! У словоформы может быть несколько лемм.

Словарная форма

1. Для существительного именительный падеж, единственное число.
2. Для прилагательных именительный падеж, единственное число, мужской род.
3. Для глаголов инфинитив.

Лемматайзеры для русского языка

Лемматизацию и/или стемминг для русского языка могут выполнять Mystem, AOT, Sphinx, Treetagger, Stemka и др.

Методы лемматизации:

- проверка в словаре (в простейшем случае можно взять викисловарь);
- перебор вариантов применения правил до получения словарной формы, которая должна присутствовать в словаре.

Теоретическую основу во многом составляет грамматический словарь русского языка А. А. Зализняка.

Другие вспомогательные задачи

1. Определение части речи (*part of speech*, *POS*).
2. Построение дерева зависимости слов в предложении (обычно начиная с подлежащего).
3. Морфологическое аннотирование (тегирование).

Спасибо за внимание!

Веб-сайт: https://marigostra.ru/
E-mail: msp@luwrain.org