Национальный исследовательский университет
Высшая школа экономики
Факультет компьютерных наук

Отчет по домашнему заданию №1

Графематический и морфологический анализ текста

студент 1 курса магистратуры
Зуев Кирилл Александрович

Москва, 2019

Содержание

1 Постановка задачи выбранного варианта (F)

Провести исследование качества разрешения морфологической омонимии одного из морфоанализаторов для русского языка, подключив его к своей программе. Исследование можно провести вручную, взяв нескольких текстов небольшого размера (20 – 25 предложений), либо автоматически, используя как эталон размеченные тексты. В последнем случае следует вычислить точность разрешения омонимии по леммам, части речи, а также по всем морфологическим характеристикам/тегам. Для исследования можно взять одну из следующих пар анализатор – размеченный текст:

- mystem – тексты НКРЯ (RNC);
- pymorphy – OpenCorpora;
- CrossMorphy – OpenCorpora.

2 Уточнение постановки задачи

Мной было выбрано **автоматическое** тестирование морфологического анализатора **pymorphy**, а в качестве эталона используются размеченные тексты со снятой омонимией из корпуса **OpenCorpora**.

3 Определение точности разрешения разных типов омонимии

Необходимо определить точность разрешения морфологическим анализатором разных типов омонимии:

- по лемме;
- по части речи;
- по всем морфологическим характеристикам.

pymorphy при обработке входного слова выдает список всевозможных вариантов его разбора по морфологическим характеристикам. В данном случае используется бесконтекстный способ разрешения омонимии, так как каждому варианту сопоставлена некоторая вероятность. Таким образом, выбирается вариант разбора с наибольшей вероятностью, то есть первый из полученного списка.

Для проверки правильности разрешения омонимии по лемме необходимо сравнить полученную лемму из набора характеристик с леммой этого же слова из размеченного корпуса. Если они совпадают, то мы считаем разрешение омонимии верным для данного типа.

В случае разрешения омонимии по части речи полученная в ходе разбора анализатором часть речи с частью речи из эталонного корпуса. И если они совпадают, то считаем, что омонимии данного типа нет и разбор слова произошел верно.

Разрешение омонимии по всем морфологическим характеристикам предполагает, что все характеристики, включая лемму и часть речи, должны быть идентичными у полученного разбора и у эталона. Если все совпадает, то считаем для данного типа омонимии разбор верным.

В результате работы программы были получены следующие результаты точности:

- точность разрешения омонимии по лемме: 0.92296;
- точность разрешения омонимии по части речи: 0.95848;
- точность разрешения омонимии по всем морфологическим характеристикам: 0.70885.

4 Код программы

Реализация программы находится в приложенных файлах **HW1.pdf** и **HW1.ipynb**.

5 Выводы по исследованию

Бесконтекстное разрешение омонимии, встроенное в морфологический анализатор **pymorphy** с большой точностью определяет верную лемму и часть речи и с достаточно неплохой точностью, но хуже, определяет все морфологические характеристики целиком.