УДК 004.822

# «Методы лемматизации при обработке текстов на русском языке»

Автор: Хабаров Ю.А., Университет ИТМО, Санкт-Петербург
Соавтор: Николаев В.В., Университет ИТМО, Санкт-Петербург
Научный руководитель: Николаев В.В., Университет ИТМО, Санкт-Петербург

В связи со стремительным ростом количества информации в электронном виде, все острее встает вопрос обработки больших массивов данных в виде текстов на естественном языке. Одним из первых этапов обработки текста на естественном языке является приведение слов в предложениях к начальной (словарной) форме. Такое приведение называется лемматизацией (нормализацией).

У процесса лемматизации можно найти несколько применений. Прежде всего, лемматизацию используют поисковые системы. Она помогает им ускорить индексирование и обработку запросов, а также повысить релевантность своей выдачи. Поисковики пропускают каждую страницу через алгоритм-лемматизатор, чтобы сохранить ее в базе в компактной и удобной для поиска форме. Запросы тоже проходят через лемматизацию. Неважно, что ввел пользователь: «куплю машину» или «купить машину» — поисковик преобразует слова в леммы («купить машина») и покажет один и тот же результат.
Другое применение лемматизации — проверка уникальности.
Кроме того, можно привести и куда более простой случай из жизни, где встречается лемматизация — достаточно посмотреть в алфавитный перечень в конце большой книжки. Приведение к простой форме позволяет быстро ссылаться в разные места, где действительно упоминается что-либо по данной теме.

Лемматизация по сути своей более сложный подход к поиску основы слова, нежели стемминг. Чтобы понять, как работает лемматизация, нужно знать, как создаются различные формы слова. Большинство слов изменяется, когда они используются в различных грамматических формах. Конец слова заменяется на грамматическое окончание, и это приводит к новой форме исходного слова. Лемматизация выполняет обратное преобразование: заменяет грамматическое окончание суффиксом или окончанием начальной формы.

Существует несколько реализаций, которые позволяют искать основу слова для текстов:

- Стеммер Портера

Основная идея стеммера Портера заключается в том, что существует ограниченное количество словообразующих суффиксов, и стемминг слова происходит без использования каких-либо баз основ: только множество существующих суффиксов и вручную заданные правила.

Алгоритм состоит из пяти шагов. На каждом шаге отсекается словообразующий суффикс и оставшаяся часть проверяется на соответствие правилам (например, для русских слов основа должна содержать не менее одной гласной).
Если полученное слово удовлетворяет правилам, происходит переход на следующий шаг. Если нет — алгоритм выбирает другой суффикс для отсечения. На первом шаге отсекается максимальный формообразующий суффикс, на втором — буква «и», на третьем — словообразующий суффикс, на четвертом — суффиксы превосходных форм, «ь» и одна из двух «н».

- Stemka

Основан на вероятностной модели: слова из обучающего текста разбираются анализатором на пары «последние две буквы основы» + «суффикс» и если такая пара уже присутствует в модели — увеличивается её вес, иначе она добавляется в модель. После чего полученный массив данных ранжируется по убыванию веса. Результат — набор потенциальных окончаний с условиями на предшествующие символы — инвертируется для удобства сканирования словоформ «справа налево» и представляется в виде таблицы переходов конечного автомата.

- Mystem

На первом шаге при помощи дерева суффиксов во входном слове определяются возможные границы между основой и суффиксом, после чего для каждой потенциальной основы (начиная с самой длинной) бинарным поиском по дереву основ проверяется её наличие в словаре либо нахождение наиболее близких к ней основ (мерой близости является длина общего «хвоста»). Если слово словарное — алгоритм заканчивает работу, иначе — переходит к следующему разбиению.
Если вариант основы не совпадает ни с одной из «ближайших» словарных основ, то это означает, что анализируемое слово с данным вариантом основы в словаре отсутствует. Тогда по имеющейся основе, суффиксу и модели «ближайшей» словарной основы генерируется гипотетическая модель изменения данного слова. Гипотеза запоминается, а если она уже была построена ранее — увеличивает свой вес. Если слово так и не было найдено в словаре — длина требуемого общего окончания основ уменьшается на единицу, идёт просмотр дерева на предмет новых гипотез.
Результатом работы стеммера является получившийся набор гипотез для несуществующего или одна гипотеза для словарного слова.

У перечисленных выше методов можно отметить важную деталь: они работают с данными, которые либо в них уже заложены (например дерево суффиксов в Mystem), либо они обучаются на каком-либо наборе текстов, либо (наименее предпочтительный с точки зрения точности вариант) просто делают некоторые преобразования над текстом (стеммер Портера). Однако в настоящее время различными исследовательскими группами создаются семантические сети, которые позволяют извлекать смысловые значения для слов. Таким образом, зная конкретный смысл слова в предложении, можно гораздо точнее решить задачу разрешения омонимии в контексте приведения слова к нормальной форме в определенном предложении, и, таким образом, точнее найти собственно начальную форму слова. Это и будет являться задачей нового исследования — написания модуля лемматизации для существующей и разрабатываемой семантической сети для текстов на русском языке.