

Яцко Вячеслав Александрович
д. филол. наук, профессор ХГУ им. Н.Ф.Катанова,
E-mail: iatsko@gmail.com
г. Абакан, РФ

ОСОБЕННОСТИ РАЗРАБОТКИ СТЕММЕРА

Аннотация

Описывается опыт разработки гибридного стеммера, который функционирует на основе списка стемм, а списка также суффиксов и окончаний. Пользователь может редактировать списки, адаптируя стеммер под свои потребности. Стеммер локализован для русского языка

Ключевые слова

Автоматическая обработка текстов, морфологический анализ, гибридный адаптивный стеммер

В области автоматической обработки текстовых документов под стеммерами понимаются программы морфологического анализа, которые позволяют распознать основу слова (стемму), удалив суффиксы и окончания. Общепринятым является деление стеммеров на алгоритмические и словарные [1; 2]. Алгоритмические стеммеры функционируют на основе файлов данных, содержащих списки деривационных суффиксов и флексий. В процессе морфологического анализа программа выполняет сопоставление суффиксов и окончаний слов во входном тексте и в соответствующем списке, причём анализ начинается с последнего символа слова. Словарные стеммеры функционируют на основе словарей основ слов. В процессе морфологического анализа такой стеммер выполняет сопоставление основ слов во входном тексте и в соответствующем словаре, а анализ начинается с первого символа слова.

Алгоритмические стеммеры более распространены, чем словарные. Это объясняется тем, что количество суффиксов и флексий в каждом конкретном языке достаточно ограничено, и изменения на уровне морфологической структуры происходят намного медленнее, чем на лексическом уровне. Списки суффиксов и окончаний обычно включают несколько сотен терминов и не требуют постоянного обновления, в то время как словари основ могут включать сотни тысяч терминов и требуют редактирования в связи с постоянным изменением словарного состава языка. Наиболее известными стеммерами для английского языка являются алгоритмические стеммеры: стеммер Портера и ланкастерский стеммер. Для русского языка был разработан алгоритмический стеммер MyStem¹.

Нами в настоящее время разрабатывается гибридный адаптивный стеммер для русского языка в рамках проекта по созданию системы автоматического анализа мнений пользователей о коммерческих продуктах². Стеммер функционирует на основе базы данных, включающих три файла: файл со списком стемм наиболее частотных слов русского языка, файл со списком суффиксов и окончаний, файл со списком стоп слов. Все три списка могут редактироваться пользователем, что позволяет адаптировать стеммер под цели конкретного проекта.

Таблица 1

Фрагмент лингвистической базы данных стеммера

Список суффиксов и окончаний	Список стемм	Список стоп слов
а авый аемый ак ака	абсолют август авиаци автобус автомат	а е и ж м

¹ <https://tech.yandex.ru/mystem/>

² Проект поддержан грантом РФФИ 16-07-00014

Обобщенный алгоритм функционирования стеммера включает следующие процедуры. Вначале в тексте фильтруются стоп слова, которые не стеммируются и выводятся в результат в неизменяемом виде. Далее с помощью списка стемм распознается стемма входного слова. Если её не удастся распознать, программа обращается к списку суффиксов и окончаний. Если не удастся распознать стемму и с помощью этого списка, слово входного текста выводится в результат. Распознавание стемм входного текста на основе списка стемм идет по направлению от входного токена к основе в списке. Выполняется посимвольное сравнение, начиная с первого символа. Если находится совпадение со стеммой, то продолжается поиск совпадения следующего символа в словах, начинающихся на ту же букву. Например, входной токен *акцентировал*. В списке есть стеммы *акц* и *акцент*. Установив совпадение первых трех символов входного токена со стеммой *акц*, следует продолжить искать совпадение четвертого символа в ближайших стеммах, начинающихся с той же буквы. Соответственно устанавливается совпадение четвертого символа токена с четвертым символом стеммы *акцент*, далее – совпадение пятого и шестого символов. Седьмой символ различается, ищется в соседних стеммах, и если не находится, на этом цикл заканчивается, на выход подается *акцент*.

В связи с большой вариативностью основ слов в русском языке в ряде случаев в списке стемм приводятся точные словоформы. Например, *спать=спал=сплю,спишь,спит,спим,спите,спят*. На выход подается стемма, идущая первой: *спать* (она может совпасть со словом); *спал* – это вспомогательная стемма, которой в тексте могут соответствовать слова *спал, спали, спала, спало, спалось*. Далее перечисляются точные словоформы, отделяющиеся запятыми. Словоформы, указанные через запятые после знака равенства отождествляются непосредственно с основной стеммой, в то время как словоформы, соответствующие вспомогательной стемме (*спал*) сначала отождествляются с ней, а затем с основной стеммой.

Распознавание стемм на основе списка аффиксов идет по направлению от входного токена к суффиксу или окончанию в списке. Выполняется посимвольное сравнение, начиная с последнего символа. После того, как найдено совпадение окончания токена с аффиксом, продолжается поиск совпадения следующего символа в аффиксах. Например, входной токен *национальность*. В списке аффиксов есть: *ь, сть, ость, ность, аль, ион*. Вначале устанавливается совпадение последнего символа *ь* с последним символом в пяти аффиксах, далее просматриваются вторые символы этих аффиксов, устанавливается совпадение *т* в трех аффиксах и входном слове. Затем устанавливается совпадение *с* – в трех аффиксах, *о* – в двух, *н* – в одном. Далее ищется *ь* в шестой позиции с конца. Не находится. Тогда ищется *ь* в конечной позиции, при этом уже проанализированные аффиксы не учитываются. Находится *аль*. Затем сопоставляется второй символ аффикса с седьмым символом токена, и устанавливается совпадение. Далее сопоставляется третий и восьмой символы и устанавливается совпадение. Далее девятый символ токена *н* сопоставляется с последним символом аффиксов и находится *ион*. В результате на выходе получится *нац*.

Программы морфологического анализа давно применяются в информационно-поисковых системах с целью повышения показателя полноты поиска. В последние десятилетия они используются и в системах автоматической классификации текстов, в частности в целях распознавания плагиата [3], поскольку совпадение лексического состава текстов более адекватно устанавливать по основам слов, а не по точным словоформам. Предложенный нами оригинальный алгоритм распознавания основ слов, как мы предполагаем, поможет повысить эффективность функционирования этих систем.

Список использованной литературы.

1. Яцко В. А. Методы и алгоритмы автоматического анализа текста // Научно-техническая информация. Сер. 2. - 2011. - № 9. - С. 12-19.
2. Moral C. A survey of stemming algorithms in information retrieval // IR information research. – 2014. – Vol.9. - No 1. - URL: <http://www.ldoceonline.com> (дата обращения 22.10.2016)
3. Kent C.K., Salim N. Web-based cross language plagiarism detection // Journal of computing. – 2009. – Vol.1. – Issue 1.- P. 39-43. - URL: <https://arxiv.org/ftp/arxiv/papers/0912/0912.3959.pdf> (дата обращения 22.10.2016)

© Яцко В.А., 2016