Влияние морфологического анализа на качество информационного поиска

© М.В. Губин А.Б. Морозов
Консорциум «Кодекс»
max@gubin.spb.ru amoro@kodeks.ru

Аннотация

Статья содержит результаты экспериментального исследования влияния различных подходов к обработке форм русских слов на качество информационного поиска. Большинство современных русскоязычных поисковых систем производят нормализацию (лемматизацию) слов, то есть приведение различных форм слова к одному поисковому признаку. Считается, что это позволяет заметно улучшить качество поиска. Известно несколько подходов к нормализации: с использованием алгоритмов усечения окончаний (стемминг), алгоритмов морфологического анализа на основании правил и/или словарей. В проведенных экспериментах использовались ряд общедоступных русскоязычных модулей стемминга и морфологического анализа. Для сравнения качества поиска использовалась методика РОМИП.

Введение

В большинстве естественных языков наблюдается такое явление, как морфологическая изменяемость слов[1]. Данное явление сильно выражено в русском языке, который относится к группе флективных языков со сложной системой флексий[2,3].

Информационно-поисковая система, работающая с русским языком, должна учитывать эту особенность языка, что реализуется обычно с помощью специального модуля системы, называемого модулем морфологического анализа. В данной работе исследуется влияние работы данного модуля на качество информационного поиска.

Использование морфологического анализа в поисковой системе

В современной поисковой системе модуль морфологического анализа обычно выполняет преобразование множества всех слов языка во множество лемм – нормализованных форм слов[4]. В литературе данный модуль поисковой системы называют модулем морфологического анализа, нормализатором слов, лемматизатором или стеммером (stemmer). Операцию, выполняемую данным модулем можно представить как отображение:

$$W \rightarrow L$$

,где

W – множество всех терминов;

L – множество всех лемм.

При этом количество лемм меньше мощности множества всех терминов $|W|>|L|$.

Реализуя данное преобразование, разработчики поисковой системы пытаются достичь следующих целей:

1. Увеличение полноты поиска. Так как отбираются документы, которые содержат все формы слова, то в результат поиска попадают не только документы со словом в совпадающей с запросом форме, но и другие документы, содержащие различные формы данного слова;
2. Улучшение точности поиска. При использовании статистических алгоритмов поиска и отбора в результат поиска нескольких документов, которые получили наибольший вес, очень важно становится получение частотных характеристик документов. При этом использование вместо частот слов частоты лемм может позволить получить больший вес для релевантных документов и тем самым поместить их во множество отобранных;
3. Упрощение пользовательского интерфейса. Так как для пользователя частой задачей является найти «все варианты упоминания», то в случае отсутствия автоматического расширения слов его вариантами пользователь вынужден изучать и использовать формулы или операторы отсечения;
4. Уменьшение размера индексной информации и ускорение обработки запроса. Так как количество лемм меньше количества слов, то лемматизация приводит к

уменьшению размера индекса и увеличению скорости обработки запроса. Однако надо сказать, что развитие методов сжатия индексной информации и совершенствование алгоритмов поиска приводит к тому, что использование морфологической обработки для уменьшения размера индекса в настоящее время не является критически важным в большинстве случаев.

Существует большое количество работ, посвященных влиянию использования различных алгоритмов морфологического анализа на качество информационного поиска в коллекциях на различных языках[5,6,7,8,9,10,11,12,13].

В данном исследовании нас интересовало, насколько доступные в настоящее время модули морфологического анализа позволяют достичь первой и второй цели.

Принципы построения модуля морфологического анализа

В настоящее время можно выделить два основных принципа построения модулей морфологического анализа для русского языка в поисковых системах:

1. На основании правил (процедурный подход). При таком построении модуль содержит набор правил морфологических преобразований. Для русского языка в основном это таблицы суффиксов и условий их отсечения, с помощью которых данный термин преобразуется к некоторой нормальной форме;
2. На основе словаря (декларативный подход). При этом преобразование термина в лемму производится с помощью специальной таблицы (словаря), которая содержит отображение множества терминов на множество лемм.

Сложность естественного языка приводит к тому, что ни один из описанных подходов не может «охватить» его целиком. Для русского языка известно более 1000 правил словообразования с множеством исключений, что делает создание полного набора правил крайне сложным. Постоянное развитие языков и большой размер словарей делает невозможным «чистое» использование декларативного подхода. Поэтому в большинстве систем используют словарь и набор правил для обработки слов, не содержащихся в словаре[14,15,16,17].

Доступные средства работы с русским языком

В настоящее время существует несколько разработок, которые можно использовать для создания модуля морфологической обработки в системах информационного поиска. Некоторые из них являются коммерческими системами, другие доступны для свободного использования. В настоящей работе мы ориентировались в основном на последние. Далее в статье приводится краткое описание использованных средств морфологической обработки русскоязычных текстов.

ISpell

Ispell[18] является распространенным opensource средством контроля правописания, однако структура словаря, содержащего основу и указание на аффикс-файл, позволяет использовать его как средство выделения нормальной формы слова.

Для использования в качестве морфологического модуля поисковой системы требуется либо осуществлять поиск в исходных файлах, либо на их основе сгенерировать базу словоформ. Файл аффиксов можно использовать как множество суффиксов для грубого морфологического анализа по правилам. Ispell является широко распространенным выбором при реализации русской морфологии во многих популярных поисковых машинах, таких как DataparkSearch[19], ht://Dig[20] и т.д.

К сожалению, по техническим причинам, экспериментальное исследование данного модуля не было проведено.

Snowball

Snowball[21] - миниязык обработки строк, предназначенный для создания стеммеров. Вместе с ним доступны готовые стеммеры для английского, французского, испанского, португальского, итальянского, немецкого, датского, шведского, норвежского, финского и русского языков. Правила стемминга описываются на языке Snowball, транслируются в код на C или Java и далее используются как обычная библиотека. Разработкой Snowball руководит Martin Porter, разработчик одного из первых подобных алгоритмов для английского языка[22].

Стеммер, созданный snowball'ом, представляет собой исходный код модуля на языке C или Java, реализующий правила стемминга для какого-либо языка. Для удобного использования полученного модуля предназначена библиотека, которая, будучи собранной, может включать в себя несколько стеммеров.

В комплект поставки входят исходные тексты библиотек для языков C и Java, исходные тексты правил стемминга для упомянутых выше языков, исходный текст демонстрационной программы.

Stemka

Stemka[23] – библиотека морфологического анализа, созданная Андреем Коваленко. Она

поставляется в исходных текстах и может свободно использоваться при ссылке на автора.

Алгоритм стемминга представлен в виде таблицы переходов конечного автомата. Каждое правило - это окончание и две последних буквы неизменяемой основы. Правила были порождены автоматически следующим образом: некоторое количество полнотекстовой информации было разбито на слова, которые подавались на вход морфологического анализатора основанного на словаре. По утверждению автора анализатор давно и успешно используется, его словарь достаточно хорошо выверен. Для всех распознанных анализатором словоформ выделялись их точная основа, отделенное окончание с двумя последними буквами основы либо формировалось новое правило, либо увеличивался вес существующего правила. Затем правила были проранжированы по вероятности встречи в текстах, и маловероятные (с вероятностью менее одной десятитысячной) были отброшены. К алгоритму также было добавлено специальное правило, гласящее, что неизменяемая основа должна содержать как минимум одну гласную.

Имеются правила для русского и украинского языков. В комплект поставки входит программа генерации правил на основе словарей ISpell.

Результат работы алгоритма - все допустимые варианты выделения формальной основы поданного на вход слова. Выбор конкретного варианта оставлен пользовательской программе.

AOT

Рабочая группа AOT разрабатывает программное обеспечение в области автоматической обработки текста[24].

С 2002 года пакет имеет лицензию LGPL для варианта в виде исходных текстов, который можно свободно использовать в некоммерческих и коммерческих проектах. Бинарный вариант пакета морфологического анализа для Windows доступен как демонстрационный продукт и содержит только русский словарь.

В системе используется морфологический словарь (имеются словари для русского, английского и немецкого языков). Словарь может содержать информацию о словах, возможных окончаниях, приставках и ударениях. Если слово не удается найти в словаре, задействуется механизм морфологического предсказания. Последний работает следующим образом: вначале делается попытка найти в словаре словоформу, которая совпадает с максимально длинной правой частью анализируемого слова (как минимум, из 4 символов, при этом размер неузнанной, левой части не должен превышать 5 символов) и если это удается, то слово предсказывается по найденной правой части; если попытка не удалась, производится предсказание по окончаниям, в результате которого выдается список вариантов слов частей речи, которые могут быть продуктивными (для русского - существительное, глагол, прилагательное и наречие) с наиболее похожими окончаниями, причем, для каждой части речи среди всех вариантов выбирается один, с наиболее часто встречающимся в словаре набором окончаний. Авторы заявляют, что скорость предсказания в 2 раза ниже скорости обычного поиска, а точность предсказания составляет 87%.

В основе поставляемого AOT русского словаря лежит словарь Зализняка.

Оценка качества морфологического анализа

В настоящее время в литературе в основном используется подход к оценке качества морфологического анализа на основании количества форм слов, неправильно отнесенных к леммам[24,25]. При этом выделяют следующие виды ошибок лемматизации:

1. Under-stemming, когда морфологические формы одного слова относят к разным леммам;
2. Over-stemming, когда разные слова ошибочно относят к одной лемме.

Для оценки этих видов ошибок, вводят две метрики UI – under-stemming index - процент терминов, для которых данный модуль лемматизации совершил ошибку under-stemming и OI – over-stemming index – процент слов, для которых морфологический модуль совершил ошибку over-stemming.

Оценка этих характеристик на практике достаточно затруднительна, так как требует наличия некоего образцового лемматизатора, который бы обеспечивал идеальное разбиение. Очевидно, что такого образцового разбиения для русского языка не существует.

В качестве другой характеристики, которую значительно проще оценить, является среднее количество слов OW, отнесенных к одной лемме данным лемматизатором. Данную характеристику можно связать с метриками UI и OI. Очевидно, что при прочих равных условиях модуль, имеющий большее значение OW, показывает большее значение OI и меньшее UI, и наоборот.

В данной работе использовался подход, когда исследовалось влияние лемматизатора при неизменных других параметрах на качество информационного поиска, которое демонстрирует система. При этом оценивалось изменение таких характеристик качества поиска как полнота (recall) и точность (precision) в зависимости от OW. Данный подход имеет следующие достоинства:

1. Используются не косвенные характеристики, а непосредственно связанные с качеством информационного поиска;
2. Все характеристики и оценки получаются по хорошо себя зарекомендовавшим и простым

методикам, что обеспечивает их простую интерпретацию и повторяемость.

К недостаткам используемого метода можно отнести то, что качество работы морфологических алгоритмов проверяется на относительно небольшой выборке слов, участвовавших в тестируемых запросах, и для заданной коллекции документов. При этом возможны ошибки, связанные с особенностями определенного морфологического модуля для данной конкретной группы слов или документов.

Постановка эксперимента

Для проведения экспериментов использовалась методика оценки качества поиска РОМИП[27]. Использовались данные конференции за 2005 год для экспериментов: коллекция narod.ru и коллекция legal. Использовались запросы, для которых были известны таблицы релевантности, что позволило полностью автоматизировать процесс выполнения экспериментов. Для каждой коллекции выполнялись эксперименты следующих видов:

1. С отключенной морфологией. При этих экспериментах не производилось никакой морфологической обработки, термины искались точно так, как они были заданы в запросе.
2. Ручное усечение. Для выполнения данных экспериментов запросы вручную модифицировались – термины запроса усекались до основ. Например, запрос rb15631: “как накачать мышцы” формулировался как “как накач* мышц*”.
3. С использованием библиотеки snowball в варианте для русского языка. Модуль stemmera использовался для получения псевдооснов, при этом к одному и тому же классу относились термины, для которых формировались одинаковые псевдоосновы.
4. С использованием стеммера Stemka. Использовался вариант модуля с набором правил, поставляемый разработчиком.
5. С использованием модулей морфологического анализа AOT. При этом к одному и тому же классу относились термины, которые отнесены модулем морфологического анализа к одной лемме.

В качестве поисковой системы использовалась поисковая машина системы Кодекс, реализующая вариант алгоритма TFIDF для формирования весов документов в зависимости от терминов и Local PageRank для учета гипертекстовых связей между документами[28,29].

Для оценки среднего значения OW использовалась следующая методика:

Был сформирован словарь всех слов, встречающихся в коллекциях РОМИП, использованных для экспериментов. Каждый из исследуемых вариантов морфологического анализа использовался для отбора из этого словаря форм слов, использованных в запросах. OW вычислялся как отношение отобранных слов к количеству форм слов взятых из запросов.

В результате экспериментов вычислялись по методике РОМИП средняя полнота, точность и строился 11 точечный график.

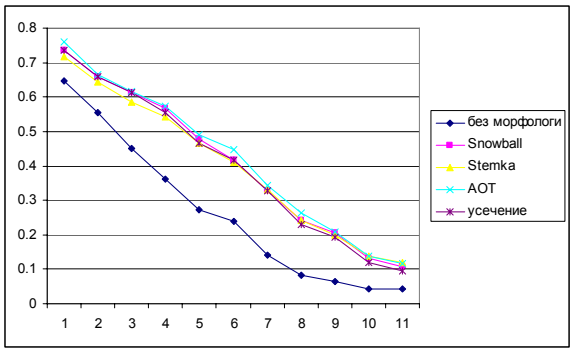

Рисунок 1. Коллекция legal, строгая релевантность

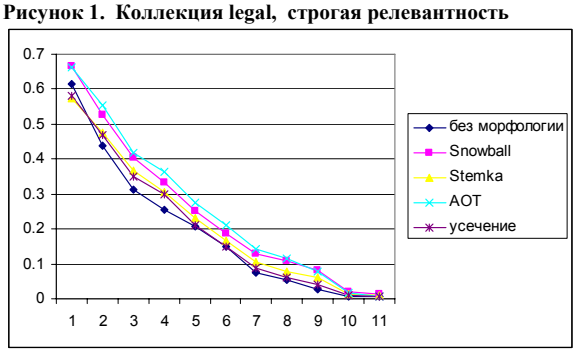

Рисунок 2. Коллекция Web, слабая релевантность

На рисунке 1 приведен пример построенных графиков для коллекции legal и строгих требований к релевантности, на рисунке 2 аналогичные графики для Web коллекции, слабых требованиях к релевантности. Графики для других требований к релевантности имеют аналогичный характер и не приводятся. Из рисунков видно, что все используемые методы учета морфологии показали очень близкие результаты. Система с отключенной морфологией показала во всех экспериментах более низкое качество поиска, причем это значительно более заметно при строгих требованиях к релевантности.

Среднее значение показателя OW для двух исследуемых коллекций приведена в Таблице 1.

Интересно, что использование модулей морфологического анализа приводило к увеличению как точности, так и полноты. Это не соответствует распространенному мнению о том, что отказ от морфологии позволяет повысить точность поиска.

Модули AOT практически для всех терминов использовали словарную базу, только 2 слова из

массива запросов legal имели опечатки, и для них не было статей в словаре. Поэтому данный алгоритм можно считать образцовым стеммером, производящим наиболее правильное отнесение форм к одной лемме с точки зрения правил русского языка. Большие значения OW для других модулей говорят о том, что данные алгоритмы имеют тенденцию к over-stemming. Причем наименее к этому склонна библиотека Snowball. Однако, как показывают значения показателей точности (Pr) и полноты (Rc) они в большинстве случаев не отличаются значительно от словарного подхода, а иногда и достигают лучших результатов. Подход с ручным усечением всегда показывал несколько худшие результаты, особенно для показателей точности, что достаточно очевидно.

Таблица 1. Показатели различных алгоритмов

Алгоритм	OW	Legal (And)		Web(OR)	
		Pr	Rc	Pr	Rc
Без морфологии	1	0.17	0.41	0.23	0.30
Snowball	9.77	0.22	0.64	0.24	0.41
Stemka	12.81	0.21	0.66	0.21	0.37
AOT	7.58	0.22	0.67	0.25	0.43
Усечение	48.1	0.18	0.57	0.17	0.35

Выводы

На основании проведенных экспериментов можно сделать следующие выводы:

1. Использование модуля морфологического анализа позволяет увеличить не только полноту, но и точность информационного поиска. Данный результат можно объяснить тем, что в случае отсутствия морфологической обработки, часто встречается ситуация, когда в выборку попадают документы, не релевантные запросу, но содержащие совпадающие формы, в то время как в релевантных документах данные слова употребляются в другой форме.
2. Введенный показатель OW (среднее количество слов OW, отнесенных к одной лемме данным лемматизатором) может быть использован для оценки ожидаемого качества поиска с использованием оцениваемого модуля морфологического анализа, так как в экспериментах для систем с близким значением данного показателя были достигнуты близкие характеристики качества поиска.
3. Исследуемые варианты подхода к реализации морфологической обработки показали похожий характер зависимостей для различных коллекций РОМИП, что говорит о том, что полученный результат распространяется как на документы «официального характера» коллекции legal, так и для «разнохарактерных» интернет-документов коллекции narod.ru.

Литература

[1] А.Ю. МУСОРИН Основы науки о языке. Новосибирск, Новосибирское книжное издательство, 2004

[2] Зализняк А.А. Грамматический словарь русского языка. Москва, Русский язык, 1980

[3] Гельбух А.Ф., Сидоров Г.О. К вопросу об автоматическом морфологическом анализе флективных языков. Труды Конференции Диалог-2005, стр. 92

[4] Frakes, W.B. & Baeza-Yates, R (1992) *Information Retrieval: Data Structures and Algorithms*, Englewood Cliffs, NJ, Prentice Hall

[5] Kraaij, W. & Pohlmann, R., 1996: "Viewing stemming as recall enhancement," in H-P. Frei, D. Harman, P. Schauble & R. Wilkinson (eds.), Proceedings of the 17th ACM SIGIR conference held at Zurich, August 18-22, pp.40-48

[6] Popovic, M. and Willett, P., (1992) *The effectiveness of stemming for natural language access to Slovene textual data,* Journal of the American Society for Information Science, 43(5), 384-390

[7] Hull, D.A. & Grefenstette, G. (1996) A Detailed Analysis of English Stemming Algorithms, Xerox Technical Report

[8] M. Kantrowitz, B. Mohit, and V. Mittal. Stemming and Its Effects on TFIDF Ranking. In Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 357--359, Athens, Greece, 2000. ACM Press. 47

[9] Carlos G. Figuerola, Raquel Gomez, Angel F. Zazo Rodrguez, and Jose Luis Alonso Berrocal. Stemming in Spanish: A first approach to its impact on information retrieval. In Carol Peters, editor, Working notes for the CLEF 2001 workshop, Darmstadt, Germany, September 2001

[10] Nazief, B. A. A. & Adriani, M. (1996), Confixstripping: Approach to Stemming Algorithm for Bahasa Indonesia. Internal publication, Faculty of Computer Science, University of Indonesia, Depok, Jakarta

[11] Divita G, Browne A, Rindflesch TC. Evaluating lexical variant generation to improve information retrieval. Presented at: AMIA Annual Symposium; 1998; Philadelphia, Penn

[12] М.С. Агеев, Б.В. Добров, Н.В. Лукашевич, А.В. Сидоров, С.В. Штернов «Отправная точка» для дорожки по поиску в РОМИП. Труды РОМИП 2003, Санкт-Петербург, стр.87

[13] Плешко В.В. Поиск с учетом словоформ русского языка Oracle Magazine, Июнь/Июль 2003
[14] Коваленко А. Вероятностный морфологический анализатор русского и украинского языков «Системный администратор» № 1, Октябрь 2002
[15] Ilya Segalovich A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. MLMTA-2003. Las Vegas, 2003
[16] Илья Сегалович, Михаил Маслов Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов Казань, ООО "Хэтер", 1998. Т. 2. С. 547-552, сентябрь 1998
[17] Аношкина Ж.Г. Морфологический процессор русского языка //Альманах "Говор", Сыктывкар, 1995, с.17-23.
[18] Ispell - GNU Project - Free Software Foundation http://www.gnu.org/software/ispell/ispell.html
[19] Поисковый движок DataparkSearch http:// www.dataparksearch.org/
[20] ht://Dig http://www.htdig.org/
[21] Snowball http://snowball.tartarus.org
[22] M.F. Porter, 1980, An algorithm for suffix stripping, Program, 14(3) pp 130−137
[23] Английский, украинский и русский морфологический анализ и анализаторы. "Машинная морфология" для любой платформы. http://linguist.nm.ru/
[24] Автоматическая Обработка Текста http://www.aot.ru/
[25] Paice, C.D. (1996) *Method for Evaluation of Stemming Algorithms based on Error Counting*, JASIS, 47(8): 632-649
[26] Sanchez, R. M (2005) *A Generalization of the Method for Evaluation of Stemming Algorithms Based on Error Counting*, Proceedings of SPIRE 2005
[27] Российский семинар по Оценке Методов Информационного Поиска http://romip.narod.ru
[28] М.В. Губин Опыт участия ИС «Кодекс» в РОМИП 2003, Труды РОМИП'2003, стр. 31
[29] М.В. Губин Участие ИПС «Кодекс» в семинаре РОМИП 2004, Труды РОМИП'2004, стр. 28

Effect of Morphological Analysis on Quality of Information Retrieval.

Maxum Gubin, Alexander Morozov

The article contains results of experimental study of an effect of different morphological analysis methods on quality of informational retrieval system search. It is commonly accepted that morphological analysis can significantly improve recall of information search. Especially this is important for inflective languages such as Russian language. We studied three public available modules which can be used for stemming or normalization. One of them uses dictionary based approach other two are rules-based. We use a system without any morphological processing as a basic line. Also we studied variant with manual suffix stripping using wildcards.

The RIRES search quality evaluation methodology was used. This methodology is similar to TREC methods and uses Russian document and query collections.

Our experiments show that morphological analysis significantly improves not only recall but also search precision. Very close characteristics of search quality was shown by all methods of morphological analysis and manual wildcards stripping.