

Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов.

Илья Сегалович, Михаил Маслов

e-mail: iseg@comptek.ru, maslov@comptek.ru

Dictionary-based algorithm of Russian morphological analysis and synthesis with suggest for morphological description of words not contained in the dictionary is described. The algorithm is used for building of texts corpora concordance, which is a part of full-text retrieval search engine with Russian morphology in products of series “Russian Classical Literature on CD-ROM™” – “Griboedov”, “Inform-Normative”.

Описываемый алгоритм морфологического анализа является словарным. Он использует электронный морфологический словарь, являющийся разработкой лаборатории компьютерной лингвистики ИППИ РАН [2]. Объем словаря – около 90 тыс. лексем (120 тыс. словооснов). Внешний формат статей электронного морфологического словаря – формат системы ЭТАП [2]. Программная оболочка электронного морфологического словаря – Яndex-Dict – разработана подразделением «Аркадия» фирмы CompTek International. Алгоритм является частью этой программной оболочки.

Основной особенностью алгоритма является то, что для словоформы, не описанной в словаре, алгоритм генерирует её предположительную модель словоизменения – одну или несколько статей в формате системы ЭТАП. Поскольку Яndex-Dict позволяет пополнять словарь новыми лексемами, сгенерированные гипотетические статьи можно сохранить в этом словаре (или другом словаре такого же типа) для дальнейшего использования.

Основные компоненты алгоритма:

- 1) Список уникальных ключей вида {*основа, идентификатор лексемы, номер основы в лексеме*} по всем основам всех лексем словаря. Список упорядочен по основам в инверсионном алфавитном порядке, поэтому основы слов с одинаковым или сходным типом словоизменения в этом списке, как правило, находятся рядом (аналогично для нормальных форм слов: [1], стр 4). Лексемы могут иметь одну или несколько основ. Так, лексемы с чередованием в основе имеют, как правило, две основы (замок – замок, замк), супплетивные - несколько основ (идти – ид, ш, ше, шед).
- 2) Унифицированный список окончаний всех лексем словаря в инверсионном алфавитном порядке.
- 3) Алгоритм морфологического анализа словоформы по отдельной лексеме (текст лексемы – во внутреннем формате Яndex-Dict). Его параметры: лексема, основа и окончание анализируемой формы. Он возвращает количество вариантов разбора и наборы грамматических характеристик по каждому варианту разбора. Если словоформа не соответствует лексеме, то возвращаемое значение количества вариантов разбора равно нулю.

Описание алгоритма.

- 1) Находятся все варианты основ — от анализируемой словоформы отрезаются все варианты окончаний.
- 2) Для каждого варианта основы, начиная с самого длинного, осуществляется бинарный поиск в инверсионном списке основ. Если вариант основы в этом списке отсутствует, то таким образом находятся «наиболее близкие» словарные основы – имеющие максимальный по длине общий «хвост». Позиция первой «наиболее близкой» основы и мера ее сходства – число совпавших символов в основе и длина окончания – запоминаются.

3) По всем вариантам основ производится следующее:

Для всех лексем, имеющих одинаковую меру сходства (одинаковую длину общего «хвоста» основы) осуществляется морфологический анализ по лексеме.

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. В этом случае по варианту основы, окончанию и лексеме, соответствующей «ближайшей» словарной основе, генерируется гипотетическая лексема – модель словоизменения для этого неизвестного слова. В случае успешной генерации эта гипотеза подается на вход морфологическому анализатору по лексеме.

Успешные варианты разбора запоминаются виде: *{Лексема(текст статьи), Варианты разбора}*.

Если результат является гипотезой и при этом такая же гипотеза уже есть, то она не запоминается повторно; вместо этого увеличивается счетчик «продуктивности» этой гипотезы.

Если среди лексем с одинаковой текущей мерой сходства есть хотя бы один вариант разбора, то переход пункту 5 с успешным результатом. Если вариантов разбора нет, то длина требуемого общего «хвоста» основы уменьшается. Если после этого длина требуемого общего «хвоста» основы стала меньше двух, то переход пункту 5 с отказом; иначе – переход к пункту 3.

4) Производится унификация гипотез по парадигмам (поскольку формат допускает неоднозначное описание парадигмы) и их фильтрация по «продуктивности» – если «продуктивность» гипотезы меньше максимальной «продуктивности» в 5 раз, то гипотеза отсеивается.

5) Конец.

Использование генерации моделей словоизменения при автоматизированном построении конкордансов корпусов текстов.

Описанный алгоритм используется при построении полнотекстовых поисковых систем с учетом морфологии русского языка в продуктах серии электронных научных изданий (ЭНИ), разрабатываемых НТЦ «Информрегистр», ИМЛИ им. М.Горького РАН и фирмой ComrTek International. К настоящему моменту выпущен один продукт из этой серии – «Информ-Норматив» – и готовится к выпуску второй – «Грибоедов».

«Информ-Норматив» содержит тексты стандартов (ГОСТы, ANSI, ISO, ИЕС, ССИТТ, ЕСМА), правовых документов, классификаторов (ОКС, ГРНТИ, УДК, ОКП).

В состав «Грибоедова» войдут тексты всех произведений А.С.Грибоедова по всем наиболее авторитетным изданиям, литература о жизни и творчестве А.С.Грибоедова (более 200 текстов), библиографические описания изданий произведений А.С.Грибоедова и литературы о нем (более 6000), аннотированные указатели имен, географических названий, периодических изданий, словарь языка А.С.Грибоедова.

В ЭНИ имеется конкорданс корпуса текстов, который является частью поисковой системы. В частности, при лексическом поиске посредством конкорданса для всех слов из запроса синтезируются все словоформы; после этого «расширенный» запрос передается поисковому механизму для осуществления поиска. Очевидно, что чем полнее и точнее будет представлена лексика корпуса текстов в конкордансе, тем качественнее будет лексический поиск.

При подготовке ЭНИ конкорданс строится в три этапа.

Этап 1. По списку всех словоформ корпуса текстов строится список всех его лексем. В этом списке для каждой "словарной" словоформы запоминаются все статьи – варианты её морфологического разбора, а для каждой "несловарной" словоформы – все гипотетические статьи, генерируемые при её морфологическом анализе. При морфологическом анализе

используются два словаря – стандартный (объемом около 90 тыс. лексем) и словарь имен собственных (объемом более 1000 лексем).

«Несловарных» словоформ в рассматриваемом корпусе текстов – порядка 50% (здесь и далее приводятся статистические данные по корпусу текстов ЭНИ "Информ-Норматив"). Для 90% из них были сгенерированы гипотезы – в среднем по 3 на каждую словоформу.

Этап 2. Производится фильтрация гипотетических статей. Основным принципом фильтрации является анализ парадигм лексем по корпусу текстов. Под парадигмой лексемы по корпусу текстов (ПКТ лексемы) здесь понимается список словоформ лексемы, которые встречаются в этом корпусе текстов.

Используются следующие эвристические правила анализа ПКТ гипотез:

1) «Включение». Если ПКТ одной гипотезы является собственным подмножеством ПКТ какой-либо другой гипотезы, то данная гипотеза отсеивается.

Вероятность «срабатывания» этой эвристики ~ 5%.

Вероятность правильной работы этой эвристики – практически 100%.

2) «Отсутствие нормальной формы». Если ПКТ гипотезы не имеет её нормальной формы, и существует гипотеза с такой же ПКТ, имеющей нормальную форму, то данная гипотеза отсеивается.

Вероятность «срабатывания» этой эвристики ~ 15%.

Вероятность правильной работы этой эвристики ~ 70 – 75%.

Этап 3. «Ручная» доработка конкорданса – удаление оставшихся неверных гипотез и добавление статей для словоформ, по которым алгоритм дал отказ. Новые статьи могут создаваться как с помощью модуля обучения морфологического словаря (алгоритм описан в [2]), рассчитанного на неспециалистов в лингвистике, так и непосредственно во входном (внешнем) формате.

Если конкорданс системы используется только для поиска, то третий этап построения конкорданса можно считать факультативным, поскольку для поисковых целей конкорданс

оказывается, как правило, вполне приемлемым. Во-первых, алгоритм генерации гипотез дает отказ в основном по неизменяемым словам (аббревиатуры, сокращения, слова типа «шоу», «малтимедиа») и для таких слов достаточно поиска по точной форме. Во-вторых, при отсутствии верных гипотез ошибочные гипотезы нередко оказываются вполне пригодными для поиска, поскольку они могут более или менее точно и полно описать список словоформ неизвестной лексемы. Пример неполного описания парадигмы: по форме «куздра» при использовании основного словаря генерируется гипотеза по образцу «мездра», имеющая только формы единственного числа. Если же гипотезы существенно ошибочны – в основном это относится к ошибочным глагольным гипотезам для существительных типа «Байкал»-«байкать» – то словоформы из парадигм этих гипотез являются «невозможными», т.е. в реальных текстах практически никогда не встречаются, поэтому добавочного поискового «шума» при использовании этих гипотез не возникает.

Программа, демонстрирующая возможности описанного алгоритма в составе Yandex-Dict, находится по адресам www.comptek.ru/alta.html и www.comptek.ru/ramb.html. Это оболочка над наиболее популярными поисковыми механизмами из тех, которые не имеют собственных средств работы с русской морфологией – зарубежной Alra Vista и отечественным Rambler-ом.

Список литературы.

1. Зализняк А. А. Грамматический словарь русского языка(словоизменение) 2-е изд., М., «Русский язык», 1980.
2. Лингвистическое обеспечение системы ЭТАП-2. М.,1989.