К ПРОБЛЕМЕ ЛЕММАТИЗАЦИИ НЕСЛОВАРНЫХ СЛОВОФОРМ[1]

TOWARD THE LEMMATIZATION OF WORD FORMS ABSENT FROM THE DICTIONARY

***Ляшевская О.Н**. (olesar@mail.ru), ВИНИТИ РАН, Москва*

В работе дается оценка алгоритма лемматизации несловарных словоформ (единиц текста, которые словарно-ориентированный морфологический анализатор не может вывести из данных исходного словаря). Представленный алгоритм устанавливает парадигматические отношения внутри массива словоформ, подбирая оптимальное членение словоформы на псевдооснову и псевдоокончание. Показано, что соединение методов простой и сложной кластеризации эффективно для морфологического пост–процессинга больших объемов текста.

1. Введение

Среди систем автоматического морфологического анализа русских текстов наибольшую точность дают анализаторы, основанные на словаре (Mystem, Dialing, Stemka и др.). Несловарные словоформы – это единицы текста, которые словарно–ориентированный анализатор не может интерпретировать стандартным способом, т. е. не может вывести из данных исходного словаря. Встречая такие формы в тексте, анализаторы как правило пытаются построить одну или несколько гипотез об исходной форме и грамматических характеристиках словоформы (см. Mikheev 1997, Сегалович&Маслов 1998, Сокирко 2004).

Как показывает статистика НКРЯ[2], несловарные элементы составляют порядка 3% общего числа словоупотреблений. Если же рассматривать словарь словоформ этого корпуса (чуть более 2 млн. единиц), то несловарные и словарные словоформы соотносятся в пропорции 45% к 55%. Присутствие несловарного слоя в таком заметном объеме создает определенные проблемы для корпусной лингвистики.

Во–первых, возможная неточность или неоднозначность определения морфологических характеристик словоформ (большинству неопознанных словоупотреблений приписывается несколько грамматических разборов) может создавать поисковый "шум" для пользователей корпусов, а также вызывать ошибки в работе синтаксических и т. п. анализаторов, учитывающих морфологические данные. Для систем снятия морфологической омонимии несловарные формы также являются камнем преткновения, поскольку 3–граммно ориентированные системы "ломаются" на 3–х несловарных лексемах, идущих подряд (ср. *Солт-лейк-сити*).

Во-вторых, весьма актуальна проблема избыточного объема хранимой информации: например, если учесть, что для одной несловарной словоформы парсер порождает в среднем 3 гипотезы морфологического разбора [Сегалович&Маслов 1998], то для корпуса современного мирового стандарта в 1 млрд. словоупотреблений это даст дополнительно 60 млн. морфологических разборов, помноженных на избыточные синтаксические, семантические и проч. теги.

Вместе с тем, массив несловарных словоформ является источником ценного лингвистического материала, а именно, новых слов и терминологии, аббревиатур, нестандартных форм склонения и спряжения. Небезынтересен и сам по себе вопрос, каков объем «словарного багажа» языка, как он меняется во времени и как соотносится со словарным запасом других языков. В настоящее время с несловарными словоупотреблениями активно работают в основном системы извлечения информации (IR), однако автоматическая обработка этого материала требуется и в лексикографии, в частности, для составления словарей новых и иностранных слов, словарей аббревиатур и имен собственных, пополнения грамматического и орфографических словарей.

Первоочередная техническая задача в этой области для языков с развитой морфологией – составление списка лексем на базе списка словоформ, или леммное сведе́ние. Между тем, даже профессиональные мультиязычные системы по составлению словарей (ср. IDM Dictionary Production System (http://www.idm.fr/), TshwaneLex (http://tshwanedje.com/tshwanelex/) и др. предполагают ручное составление словника, поддерживая лишь сортировку по началу и концу слова. Не подвергая сомнению роль человека в отборе лексики для словника и его редактировании, мы тем не менее хотели бы обсудить перспективы компьютеризации леммного сведения как необхо-

[1] Исследование выполнено при поддержке программы научных проектов «Интернет-математика 2007» ООО «Яндекс».

[2] По данным на янв. 2007 г.: около 135 млн. словоупотреблений, морфологический парсер Mystem основан на словаре в 80-90 тыс. лексем.

димого модуля в системах и выявить возможные риски в построении систем автоматического пополнения грамматического словаря (Daciuk 2001). Данная работа преследует цель оценить эффективность одного из методов леммного сведения, суть которого состоит в установлении парадигматических отношений внутри массива несловарных словоформ.

2. Парадигматическое леммное сведе́ние (кластеризация)

Работу «гипотетического» модуля большинства русских морфоанализаторов можно образно представить следующим образом. Сначала программа порождает полное множество словоформ, предсказываемых собственным словарем. Встречая в тексте словоформу, не входящую в это множество, программа сравнивает ее с близкими по окончанию словарными словоформами и приписывает ей аналогичную грамматическую информацию. В дальнейшем для оптимизации числа разборов применяются некоторые эвристики, такие как приписывание дополнительных гипотез о несклоняемой форме, удаление или понижение в ранге гипотез с редкими и непродуктивными грамматическими разборами, удаление гипотез без гласной в основе, приоритет гипотезы с самым длинным окончанием и др. (Сегалович 1998, Коваленко 2002, Segalovich 2003, Сокирко 2004, Hana&Feldman 2004).

Как видно, всякий раз программа строит гипотезы без обращения к предыдущему опыту. Например, форме *гипермаркетов* анализатор приписывает два разбора {*гипермаркет=S|гипермаркетов(ый)=A*}, «забывая», что до этого в тексте ему встретилась форма *гипермаркеты*, не имеющая адъективного разбора. Кажется очевидным, что программный модуль, анализирующий накопленный опыт гипотетических разборов, мог бы в определенной мере снизить неоднозначность морфологической аннотации.

Парадигматический подход к лемматизации по сути имитирует работу лексикографа, который «наметанным глазом» вычленяет в упорядоченном массиве группы, относящиеся к одной лемме (см. табл. 1). Лингвистическим обоснованием этого подхода является следующее допущение: если некоторое слово открытого (словоизменительного) класса встретилось в тексте в форме X, то скорее всего оно встретится в тексте в форме Y, отличной от первой (Hana&Feldman 2004). Естественно, эта закономерность будет иметь бо́льшую силу для высокочастотных слов и на больших массивах текстов.

Freq	Словоформа
657	генома
10	геномах
83	геноме
14	геномика
12	геномике
35	геномики
59	геномной
38	геномных
167	геномов
28	геномом
17	геному
27	геномы
11	генотипирование
42	генотипирования

Табл. 1. Фрагмент частотного списка несловарных форм

Некоторые эвристики на основе парадигматического подхода (ПКТ, или “парадигма лексем по корпусу текстов”) описаны в Сегалович&Маслов 1998, Segalovich 2003 для анализатора Mystem, но в текущей версии анализатора по-видимому не используются как нерелевантные для поисковых задач. Гораздо большая роль отводится этому подходу в работе Ножов 2003 (“метод подбора словоформ на одну лексему“). Здесь предлагается удалять ложные варианты разборов, используя корреляцию по гипотезам основ и значениям классифицирующих грамматических категорий (часть речи, тип словоизменения, род имени существительного). Метод парадигматического сравнения применяется также в анализаторах других флективных языков, в частности, чешского (Hana&Feldman 2004, Kanis&Müller 2005).

Процедура автоматического сведе́ния парадигм предполагает предварительное разделение словоформ на псевдооснову[3] и псевдоокончание, причем последнее должно входить в множество окончаний русского словоиз-

[3] Псевдоосновой считается совпадающая часть всех словоформ парадигмы (мо|гу, мо|жет, мо|гли), ср. объединение тематического элемента и аффиксального элемента в расширенную флексию в Бидер и др. 1978.

менения (наш список окончаний был составлен на основе Зализняк 1977/2003). Каждой словоформе сопоставляется набор вариантов такого членения: например, словоформе “гипермаркеты” приписывается набор гипотез {гипермаркеты|, гипермаркет|ы, гипермарке|ты}. Затем каждой гипотезе приписывается вес в зависимости от того, сколько раз та или иная псевдооснова встретилась в разборах разных словоформ (см. табл. 2, столбец 4).

Словоформа	Псевдооснова	Образец	WAbs
гипермаркет	гипермаркет|	ср. "паркет", "анкет", "решет", S	4
	гипермарке|т	ср. "одет", V	3
гипермаркета	гипермаркета|	ср. "вполоборота", ADV	1
	гипермаркет|а	ср. "паркет|а", "анкет|а", "решет|а", S	4
	гипермарке|та	ср. "одет|а", V	3
гипермаркетов	гипермаркетов|	ср. "фиолетов", "бертолетов", A; "гитов", S	1
	гипермаркет|ов	ср. "паркет|ов", "облак|ов", S	4
гипермаркеты	гипермаркеты|	ср. "комроты", S, "трикраты", ADV	1
	гипермаркет|ы	ср. "паркет|ы", "анкет|ы", "счет|ы", S	4
	гипермарке|ты	ср. "одет|ы", V	3

Табл. 2. Оценка гипотез членения словоформ на псевдооснову и псевдоокончание

Различаются простая и сложная кластеризация слоформ. В первом случае из морфологической аннотации несловарной словоформы удаляются (или понижаются в ранге) все разборы, у которых вес гипотезы о псевдоосновe ниже максимального (в нашем случае это псевдоосновы “гипермарке=” с абсолютным весом 3, “гипермаркета=”, “гипермаркетов=”, “гипермаркеты=” с абс. весом 1). Сложная кластеризация включает проверку совместимости всех окончаний в одной парадигме (по данным существующих парадигм морфологического словаря). Эта процедура серьезно усложняет алгоритм, но зато позволяет исключить случаи, когда к одной парадигме ошибочно приписываются словоформы двух и более реальных лексем, ср. “барион|“ и “барион|ный”; “шмон|“ и “шмон|али”; “Александровск|” и “александровск|ий”.

Далее в работе мы опишем три эксперимента, проведенных на массиве несловарных слов НКРЯ, которые показывают преимущества и недостатки простой и сложной кластеризации.

3. Эксперимент 1. Простая кластеризация потенциальных парадигм

В качестве исходных данных для наших экспериментов был взят частотный список несловарных словоформ НКРЯ, а также сопоставленный ему массив, содержащий информацию о частоте сочетаемости этих словоформ со знаками препинания (левые и правые «соседи») – точкой, дефисом и скобкой. В частотном списке была сохранена информация о капитализации слова в тексте: прописная и строчная буквы во всех позициях различались. Из частотного списка были исключены:

1) словоформы, содержащие цифры и латинские буквы (“1991“; “approx“ и т. п.);

2) потенциальные аббревиатуры: а) словоформы без гласных (“МЖК”, “мкм”, “нрзб”); б) словоформы, состоящие из смеси больших и малых букв, исключая капитализацию (“РайПО“, “ГАБТа”); в) словоформы, после которых в тексте обычно следует точка (“ул.”, “ок.”, “англ.”); г) словоформы, после которых в тексте обычно следует открывающая квадратная или угловая скобка (“прост[ого]“, “участн<ик>“);

3) части сложных слов (“лже”, “итало”), слова–окончания (“ый”, “тонный”, “ание“, ср. *1-ый, 45-тонный, изд[ание]*): словоформы, которые обычно встречаются перед дефисом, а также после дефиса или скобки;

4) потенциальные имена собственные: словоформы, вариант капитализированного написания которых превышает установленный порог (90%).

Оставшиеся словоформы составили Основной частотный список несловарных словоформ.

Для проведения первого эксперимента был создан рабочий массив, в который из Основного списка вошли словоформы с порогом частотности 0,1 ipm, всего ок. 21 тыс. словоформ.

Применение метода простой кластеризации дало следующие результаты (табл. 3)

Таким образом, покрытие составило 64% массива несловарных словоформ. Для ряда словоформ метод простой кластеризации предсказал два варианта членения основы и окончания с равным весом гипотез, ср.

13 инновацио|нный и инновацион|ный
12 поздней|ший и позднейш|ий
11 неоконч|енная и неоконченн|ая
7 госслуж|ащий и госслужащ|ий
4 аудиосист|ема и аудиосистем|а

Число словоформ в парадигме	Число парадигм	Нарастание покрытия
13 и более	11	0,75%
12	25	2,2%
11	14	2,9%
10	38	4,7%
9	48	6,8%
8	63	9,2%
7	85	12,0%
6	139	16,0%
5	263	22,3%
4	447	30,8%
3	877	43,3%
2	2197	64,2%
Итого	13485	64,2%

Табл. 3. Результаты эксперимента 1

В этом случае была применена простейшая эвристика: вес гипотезы с более короткой псевдоосновой был принудительно уменьшен.

Для того, чтобы оценить аккуратность метода простой кластеризации, мы провели выборочную ручную проверку полученных результатов: были проанализирован состав каждого десятого кластера объемом от 5 до 18 словоформ и каждого пятого кластера объемом от 2 до 4 словоформ.

В соответствии с общими принципами русского словоизменения, самые объемные кластеры – от 11 до 18 словоформ в парадигме – включали словоформы глаголов и прилагательных (*приватизировать, склеротизованный, новогородский* и др.). Кластеры с числом словоформ от 2 до 10 содержали, помимо глагольных и адъективных, парадигмы имен существительных. В кластерах с 4–мя и более словоформами было обнаружено два вида ошибок. Во–первых, это кластеры, в которых словоформы относятся к разным леммам (2%). Сюда относятся случаи объединения форм возвратного и невозвратного глаголов (*хватить* и *хватиться*, *мерять* и *меряться*), форм наречия и прилагательного (*геополитически* и *геополитический*, *клево* и *клевый*), прилагательного и однокоренного глагола (*розный* и *розниться*), форм существительных разного рода (*латин* и *латино*, *отморозок* и *отморозка*), а также случаи совпадения нестандартных вариантов словоизменения у однокоренных слов: *родна* и *родясь*; *бось* (разг. вариант формы “бойся”) и *босый, босу*; *ложить* и *ложись* (неучтенный в морфоанализаторе вариант императива от *ложиться*)[4].

Во–вторых, ошибку дала вышеупомянутая эвристика для форм причастий, которые не были предсказаны морфологическим словарем морфоанализатора: *берущий*, *кишащий*, *привыкший, повисший, остывший, руководимый*, *предводимый*, *настоянный*. Словоформы были правильно объединены в кластеры, но неправильно поделены на псевдооснову и псевдоокончание, так как личные формы глагола отсутствовали в массиве несловарных форм (из вариантов членения “берущ|ий“ и “бер|ущий“ был выбран более длинный вариант основы). По нашему мнению, ошибки второго вида не свидетельствуют о недостатках выбранного метода, поскольку зависят от реализации конкретного морфоанализатора.

В кластерах с 3–мя словоформами ошибочно было сведено 3% кластеров. Как и следовало ожидать, кластеры с 2–мя словоформами показали самый ненадежный результат, до 15% по разным выборкам (ср. *столькие* и *стольку*, *чернию* и *черну*, *баско* и *баскет*, *шоба* и *шоблы* и т. д.).

4. Эксперимент 2. Кластеризация с понижением порога

Чтобы увеличить покрытие массива кластерами, а также увеличить объем кластеров, мы повторили процедуру простой кластеризации, добавив в рабочий массив словоформы Основного списка с частотностью более 2–х словоупотреблений в корпусе (0,099 – 0,021 ipm). Цель эксперимента 2 состояла в том, чтобы найти для частотных, но некластеризованных в результате эксперимента 1 словоформ «соседей» по парадигме среди низкочастотных словоформ. В результате было кластеризовано еще 22% из 21–тысячного списка высокочастотных словоформ (> 0,1 ipm), см. табл. 4.

Итоговое распределение объема кластеров на массиве высокочастотных словоформ можно видеть на рис. 1. 49,3% кластеров содержат 4 и более словоформы, 17% кластеров – 3 словоформы, 20,1% кластеров – 2 словоформы; некластеризованными остались 2868 словоформ (13,7%).

[4] Приведем также отдельные несистемные ошибки в кластеризации: *робят* и *робить*, *страм* (‘срам’), *страт* и *стрит*, *скин* и *скинемся, сторы* (18 в.: ‘шторы’) и *стори* (ср. «лав стори»), *прешься* и *пром*.

Число словоформ в парадигме	Число парадигм	Нарастание покрытия
8 и более	105	64,7%
7	121	65,3%
6	178	66,2%
5	314	67,6%
4	689	70,9%
3	1232	76,8%
2	2004	86,3%
Итого	4643	86,3%

Табл. 4. Кластеризация одиночных словоформ с использованием массива низкочастотных форм

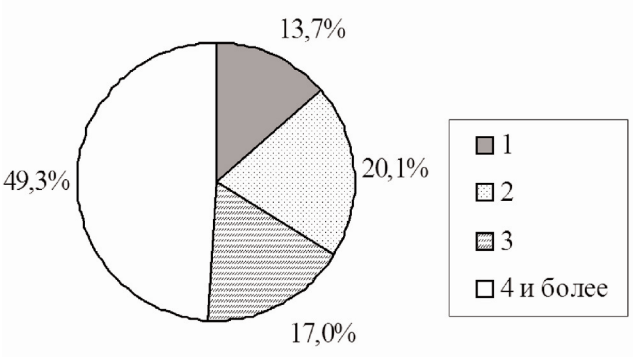

Рис. 1. Итоговое распределение объема кластеров на массиве высокочастотных словоформ

5. Эксперимент 3. Сложная кластеризация потенциальных парадигм

Для проведения сложной кластеризации мы использовали базу данных русского словоизменения, где для каждого словоизменительного типа было указано: множество псевдоокончаний (МПО), которые могут принимать формы данного типа парадигмы; часть речи и иные словоклассифицирующие характеристики леммы; инструкция для построения исходной формы (номер псевдоокончания) и, как опция, ограничения на тип основы (допустимые символы в конце псевдоосновы)[5]. Для каждого полученного ранее кластера словоформ (потенциально входящих в общую парадигму) требовалось установить, являются ли псевдоокончания совместимыми друг с другом (т.е. найти хотя бы одно МПО, для которого данное множество окончаний является подмножеством). Для совместимых окончаний создавался индекс всех МПО, в которое они входили как подмножество, и на его основе строился список возможных лемм и их разборов (элементы, совпадающие по исходной форме, частеречным и другим грамматическим характеристикам с уже внесенными в список леммами, удалялись). Кластеры с несовместимыми окончаниями помечались особым образом.

Насколько эффективно и аккуратно метод сложной кластеризации различает неправильно сведенные леммы? Метод показал свою наибольшую действенность на кластерах объемом в 2 словоформы: с его помощью было обнаружено, что 4,9% двухсловных кластеров имеют несовместимые окончания, при том что в общем для кластеров объема 2..18 этот показатель составил 1,7%. Как видно, процент обнаруженных ошибок оказался ниже, чем наша эмпирическая оценка (см. п. 3). Это связано с тем, что формы возвратных и невозвратных глаголов, наречий и прилагательных остались сведенными (ср. *наказуется* как форма страдательного залога и возвратного глагола; *геополитически* как краткая форма прилагательного, по образцу *бро́ски*, и наречия). Кроме того, не были разведены некоторые неизменяемые слова, например, кластер *завтре-завтря-завтрему*, который сравнивался с образцом *сине-синя-синему*, ср. также *куми-кумите*, *полтона-полтонны* и др.

Ручной анализ списка кластеров с несовместимыми окончаниями выявил два интересных следствия применения этого метода. Во–первых, программа «отказала» в кластеризации архаичным вариантам склонения/спряжения (*стои*, *зриши*, *есмы, есмь*). Во–вторых, были разведены парадигмы существительных на *–ие* и *–ье*, имеющие, в принципе, общую форму род. п. мн. ч. на *–ий*, ср. *думания* и *думанья*, *позвякивание* и *позвякиванье*.

6. Заключение

Мы исходили из принципа, что если некоторое слово встречается в корпусе текстов в форме f1, то весьма вероятно, что оно должно встретиться и в других формах f2, f3, …. Этот принцип, однако, не распространяется на неизменяемые слова (несклоняемые существительные и прилагательные, предлоги, союзы и другие неизменяемые части речи; наречия, которые в большинстве своем редко образуют степень сравнения). Идеальная реализация данного постулата означала бы, что в корпусе мы имели бы массив лемм, представленный 4–мя и более словоформами, с одной стороны (изменяемые леммы), и массив лемм, представленный 1 словоформой (неизменяемые леммы). На практике же значительную долю результирующего списка занимают леммы, представленные всего 2–мя словоформами, и именно они демонстрируют «слабое место» предложенного подхода (например, кла-

[5] В эксперименте 3 ограничения на тип основы не учитывались. Это позволило свести воедино нестандартные варианты словоизменения (выравнены-выравненный, лицем-лицы, болгаре-болгаров, плечей-плечьми, детем-детями, встретя-встретясь, уставя-уставясь, добродетелию-добродетельми и др.), однако привело к некоторым случайным объединениям разных лемм (см. ниже).

стер из двух словоформ на *–у* и на *–и* можно интерпретировать или как репрезентантов глагола (ср. *гну, гни*), или как два случайно объединенных неизменяемых слова (ср. *Перу, Пери*)). Принять правильное решение в этих случаях может только человек, причем если словоформы ему незнакомы, может потребоваться знание контекста.

В работе были рассмотрены две процедуры леммного сведения, позволяющие свести к минимуму объем ручного постредактирования и ранжировать массив несловарных словоформ: простая и сложная кластеризация. В результате простой кластеризации исходный массив огрублено разбивается на множества словоформ, потенциально образующих общую парадигму. Алгоритм характеризуется простотой, быстродействием, дает хорошее покрытие для частотных словоформ и, как правило, устанавливает правильное деление форм на псевдооснову и окончание. Точность алгоритма падает на 2-словных кластерах. Процедура сложной кластеризации, проверяющая найденные парадигмы на соответствие стандартным типам русского словоизменения, строит множество гипотез об исходной форме и грамматических характеристиках леммы и минимизирует ошибки в кластеризации. Вместе с тем, процедура времяемка и не всегда адекватно кластеризует неизменяемые слова и нестандартные варианты изменения словарных слов. В соединении оба подхода эффективны для морфологического пост–процессинга больших объемов текста.

Список литературы

1. Бидер И.Г., Большаков И.А., Еськова Н.А. Формальная модель русской морфологии. Ч. 1–2. М., 1978.
2. Зализняк А.А. Грамматический словарь русского языка: Словоизменение. М., 1977; 4-е изд.: М., 2003.
3. Ковалева Е.С. Особенности применения метода скрытых Марковских моделей для морфологической разметки текстов на русском языке. Дипл. работа. М.: МГУ, 2006.
4. Коваленко А. Стемка – морфологический анализ для небольших поисковых систем. Системный администратор №1, октябрь 2002.
5. Ножов И.М. Реализация автоматической синтаксической сегментации русского предложения. Дисс… канд. тех. наук. М.: РГГУ, 2003.
6. Сегалович И., Маслов М. Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // Диалог’98. Казань, 1998. Т.2.
7. Сокирко А.В. Морфологические модули на сайте www.aot.ru // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог’2004». М., 2004.
8. Segalovich I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine // MLMTA‘03, Las Vegas, NE, 2003.
9. Hana J., Feldman A. Portable language technology: Russian via Czech // Proceedings of the Midwest Computational Linguistics Colloquium, 2004, Bloomington, Indiana.
10. Daciuk J. Computer-assisted enlargement of morphological dictionaries: Finite state methods in natural language processing // Workshop at 13th ESSLLI. Helsinki, 2001.
11. J. Kanis, L. Müller. Automatic Lemmatizer Construction with Focus on OOV Words Lemmatization // Text, Speech and Dialogue 2005. Berlin/Heidelberg: Springer, 2005, 132-139.
12. Mikheev A. Automatic rule induction for unknown word guessing // Computational Linguistics, 23(3):405-423, 1997.
13. Yablonsky S. A. Russian morphological analysis // Proceedings of VEXTAL. 1999.