Автоматизация построения словаря на материале массива несловарных словоформ

О. Н. Ляшевская
ВИНИТИ РАН
olesar@mail.ru

Д. В. Сичинава
ВИНИТИ РАН
mitrius@gmail.com

Б. П. Кобрицов
ВИНИТИ РАН
neuralman@yandex.ru

Аннотация

Несловарные формы – единицы текста, отсутствующие в словаре программы морфологического анализа – представляют проблему как для автоматического парсинга текста, так и для создания словарей, основанных на текстовых корпусах. Алгоритм их лемматизации объединяет несловарные словоформы в кластеры, которым сопоставляется информация о части речи, исходной форме и других грамматических характеристиках лексемы. Процедура кластеризации включает порождение множества гипотез для каждой словоформы в соответствии с моделью русского словоизменения А. А. Зализняка и выбор в качестве наиболее вероятной той, которая чаще всего повторяется в разборах других словоформ массива.

Оценка эффективности алгоритма проводилась на материале словника Национального корпуса русского языка и набора данных «База словоформ Яндекса».

1. Введение и обзор ключевой литературы по вопросу

Большинство популярных компьютерных программ морфологического анализа русских текстов (Mystem, Dialing, Starling и др.) работают с помощью встроенного в них грамматического словаря. Узкое место таких систем – несловарные словоформы, данные о которых в словаре отсутствуют, но которые, тем не менее, все же являются словами русского языка. Среди них фамилии, имена, топонимы и другие имена собственные, термины, заимствования и прочие неологизмы, сокращения, опечатки и т. п. Механизм работы морфоанализаторов как правило предусматривает порождение ряда гипотез об исходной форме и грамматических характеристиках этих форм [1, 2, 3]. Очевидно, что точность лемматизации при этом падает.

Дунаев
Дунаев=S...;
Дунаедать=V...;
Дунай=S...
коммандос
коммандосый=A...
Ладену
Ладевать=V...;
Ладена=S...
Талибана
Талибан=S...;
Талибанный=A...;
Талибаный=A...

По статистике Национального корпуса русского языка[1], в котором несловарные элементы маркированы особым образом, эти формы составляют порядка 3 % общего числа словоупотреблений. Из 50 тыс. наиболее частотных лексем НКРЯ на долю несловарных слов приходится 16 %. В полном конкордансе словоформ НКРЯ словарные и несловарные формы соотносятся как 55 и 45 %. Аналогичная пропорция (56,4 и 44,4 %) наблюдается в частотном словаре словоформ Яндекса, размеченном тем же анализатором.

Очевидно, что несловарные слова – это важный материал как для лингвистов-теоретиков, так и для разработчиков прикладных лингвистических продуктов, составителей онтологий и IR-систем. В лексикографии эти слова служат источником для словарей неологизмов, аббревиатур, терминов, иностранных слов, географических названий, имен и фамилий, а также для пополнения уже существующих словарей. Отдельная область несловарного материала – это присутствующие в электронных текстах нестандартные формы склонения и спряжения, которые должны тем или иным образом быть учтены в орфоэпическом и грамматическом словарях.

Поскольку слова в словаре должны быть представлены в своей основной форме, то первой задачей компьютерной лексикографии для русского языка в данном случае является правильная лемматизация словоформ и сведение словоформ одной леммы воедино. Насколько нам известно, существующие профессиональные мультиязычные системы по составлению словарей (ср. IDM Dictionary Production System (http://www.idm.fr/), TshwaneLex (http://tshwanedje.com/tshwanelex/) не предлагают такой опции, как автоматическая компиляция лексем, поддерживая лишь сортировку по началу и концу слова.

Методы автоматического пополнения словника начали разрабатываться с середины 80-х годов [4, 5, 6]. В [7] описана процедура построения словаря на основе лексических правил с ручной посткоррекцией, первоначально тестировавшейся на русском материале (модель Смысл ⇔ Текст). Другой, словарно-ориентрованный подход – порождение гипотез для несловарных слов; такие модули используются во многих известных русских морфоанализаторах (см. выше). Схема их работы такова: порождается полное множество словоформ, предсказываемых собственным словарем; затем, 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

[1] По состоянию на январь 2007 г. автоматически размеченный корпус ок. 135 млн. словоупотреблений, парсер Mystem, встроенный словарь 80–90 тыс. лемм.

без гласной в основе, приоритет гипотезы с самым длинным окончанием и др. [2, 3, 8, 9, 10].

Неудобство применения «гипотетических» модулей в последующем ручном постредактировании словника состоит в том, что для одной несловарной словоформы порождается в среднем три гипотезы морфологического разбора [2], а точность объединения словоформ в леммы далека от абсолюта. Наш собственный опыт обработки несловарного материала НКРЯ в 2004 г. показал, что во многих случаях проще вручную породить исходную форму леммы, чем пытаться догадаться какое слово скрывается за неправильным разбором типа «копанить» (гипотеза леммы для словоформы «копань»).

Определенные улучшения в результаты лемматизации несловарных слов может привнести снятие морфологической омонимии с учетом контекста слова в тексте. В [11, 12] описаны алгоритмы дизамбигуации, работающие на множестве би- и триграммов тренировочного корпуса. К сожалению, специальной оценки эффективности работы программ для несловарных слов не проводилось.

В [2, 9] высказана идея учитывать при лемматизации не только данные грамматического словаря, но и опыт других разборов несловарных слов в тексте. Было введено понятие «парадигмы лексем по корпусу текстов» (ПКТ) – списка словоформ лексемы, которые встречаются в анализируемом корпусе текстов. В частности, эвристика, опирающаяся на сильное утверждение, что правильная лемма должна присутствовать в тексте в своей исходной форме, имела 15 %-ю вероятность срабатывания и правильно работала в 70–75 % случаев. В текущей версии анализатора Mystem эта эвристика не используется как нерелевантная для поисковых задач.

Значительно большая роль отводится парадигматическому подходу в работе [13] («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 [10, 14].

2. Идея исследования

В настоящей работе мы ставили цель оценить эффективность метода приведения словоформ к общей лемме для задач автоматического составления словника словаря.

Известное утверждение «Если новое слово встретилось в тексте, то скорее всего оно встретится в тексте еще раз» применительно к флективным языкам может звучать так:

«Если некоторое слово открытого (словоизменительного) класса встретилось в тексте в форме X, то скорее всего оно встретится в тексте в другом синтаксическом окружении, а следовательно, в форме Y, отличной от первой» [10].

Программа лемматизации на основе парадигматического подхода по сути имитирует работу лексикографа, который должен вычленить в упорядоченном массиве группы, относящиеся к одной лемме (табл. 1).

Таблица 1

Фрагмент частотного списка несловарных форм

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

Процедура автоматического сведения парадигм на первом этапе разделяет словоформу на псевдооснову и псевдоокончание. Псевдоосновой считается совпадающая часть всех словоформ парадигмы (мо|гу, мо|жет, мо|гли), ср. объединение тематического элемента и аффиксального элемента в расширенную флексию в [15]. Псевдоокончание должно входить в множество окончаний русского словоизменения. Каждой словоформе сопоставляется множество гипотез <псевдооснова, индекс парадигмы, окончание исходной формы>. Например, форме «генома» будут приписаны следующие гипотезы:

<геном, S1_m, ∅>
<геном, S1_n, о>
<геном, S1_f, а>
<геном, A1, ый>
<геном, A2, ой>
<геном, V5b1, ать>
<ген, V6b1, ать>

Затем по всему массиву гипотетических разборов подсчитывается число повторений каждой гипотезы; это и есть ее абсолютный вес. В результате на выходе программы имеется массив несловарных словоформ, которым сопоставлены варианты лемматизации, ранжированные по степени вероятности.

При составлении алгоритма лемматизации мы исходим из того, что словоформы принадлежат одному из трех словоизменительных классов: имя существительное, имя прилагательное, глагол. Предполагается, что местоимения и числительные являются закрытыми классами, все элементы которых уже перечислены в словаре. Наречия положительной и сравнительной степени считаются отдельными леммами. Таким образом, в идеале мы должны получить кластеры лемм имен и глаголов, представлен-

ные несколькими словоформами, и остаток массива, представленный неизменяемыми словами: наречиям, предикативами, междометиями, частицами, несклоняемым существительным и аббревиатурами.

Так ли это на самом деле? В исследовании была поставлена задача дать количественную (стандартные метрики точности и полноты) и качественную (лингвистическую) оценку состава массива, полученного в результате работы программы.

3. Наборы данных

В качестве исходных данных были использованы частотный список словоформ Национального корпуса русского языка (2 млн единиц) и набор данных «Частотный словарь» ООО Яндекс (1 млн наиболее частотных русских словоформ, получен сливанием статистики из 1/12 базы Яндекса). В обоих списках капитализированные и некапитализированные варианты написания различались.

Также использовалась матрица псевдоокончаний русского словоизменения (с учетом чередований), составленная на основе Грамматического словаря русского языка [16].

4. Процедуры и эксперименты

4.1. Вспомогательные процедуры

Предварительно частотные списки были размечены с помощью программы морфологического анализа Mystem, в результате были сформированы массив несловарных слов НКРЯ (913 тыс. единиц) и массив несловарных слов Яндекса (444 тыс. единиц).

Дополнительно к массиву несловарных слов НКРЯ были применены следующие процедуры:

а) «чистка» (удаление слов, содержащие цифры, латинские буквы и псевдографику);

б) нормализация («сливание» словоформ, написанных прописными и строчными буквами);

в) выделение массива имен собственных (словоформ, капитализированное написание которых превысило порог в 90 %);

г) выделение массива аббревиатур, куда вошли словоформы без гласных и словоформы, состоящие из смеси больших и малых букв, за вычетом капитализации;

д) выделение массива сокращений с точкой типа «ул.», «тел.»;

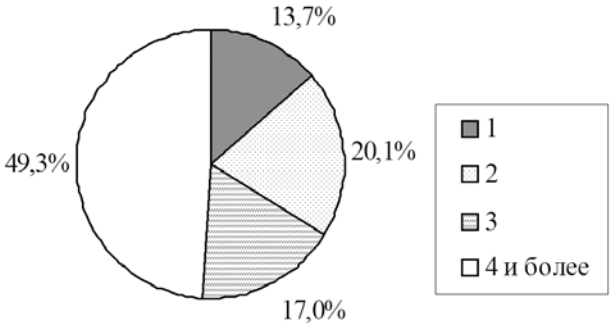
е) выделение массива префиксоидов типа «авиа-», «теле-», «псевдо-», «германо-», а также слов окончаний типа «-ый», «-ский» и др.

Основной тестовый массив НКРЯ составил нормализованный массив, из которого были вычтены элементы массивов (в-е). В дополнение к нему, на материале 6-миллионного корпуса НКРЯ со снятой грамматической омонимией был собран массив из 10 тыс. наиболее частотных несловарных словоформ. На этом массиве оценивалась точность лемматизации, так как для каждой словоформы был известен правильный вариант лемматизации, указанный разметчиками.

К массиву несловарных слов Яндекса была применена процедура нормализации. К сожалению, вследствие специфики Интернет-данных[2] попытка выделить массив имен собственных с помощью подбора порога капитализированного написания не привела к успеху.

4.2. Простая кластеризация словоформ

В статье [17] подробно описаны три эксперимента по простой и сложной кластеризации массива НКРЯ. При упрощенной кластеризации проверяется только совпадение основ словоформ, совместимости окончаний согласно стандарту парадигм русского словоизменения не требуется. Этот подход характеризуется простотой, быстродействием, дает хорошее покрытие для частотных словоформ и, как правило, устанавливает правильное деление форм на псевдооснову и окончание.



Подбор членов парадигмы для 21 тыс. наиболее частотных несловарных форм НКРЯ с помощью простой кластеризации

В ходе экспериментов была разбита на кластеры (потенциальные парадигмы) 21 тысяча наиболее частотных несловарных форм НКРЯ (>0,1 ipm), в дополнение к этому для одиночных словоформ был организован поиск «соседей» по парадигме в низкочастотной части массива. В итоге 49,3 % кластеров содержали четыре и более словоформы, 17 % кластеров – три словоформы, 20,1 % кластеров – две словоформы; 2868 словоформ остались некластеризованными (13,7 %).

Выборочная ручная проверка результатов показала, что среди кластеров объемом от 4 до 18 словоформ вероятность ошибки составила 2 %, среди кластеров объемом три словоформы – 3 %, в 2-словных кластерах вероятность ошибки резко возрастала до 15 %.

[2] Тексты в Интернете отредактированы в меньшей степени, чем тексты корпуса, в них больше опечаток и графических эффектов. Например, капитализация слова может быть обозначена не кодом заглавной буквы, а увеличением размера шрифта или с помощью рисунка. Возможно, капитализация слов была также частично потеряна при архивировании страниц. Кроме того, часть словоформ в словнике имела написание вида «яюир» вместо «САЙТ», так как была распознана поисковой системой в неправильном формате (формат W1251 вместо KOI8).

4.3. Кластеризация с проверкой типа парадигмы

При составлении словника для словаря необходимо знать исходную форму лексемы, ее часть речи и, как правило, другие грамматические характеристики. Простая кластеризация парадигм не дает возможности установить эту информацию, кроме того, в один и тот же кластер словоформ могут попадать слова разных частей речи (ср. *розный* и *розниться*). Сложная кластеризация (далее – кластеризация) включает проверку совпадения всех трех элементов гипотетического разбора <псевдооснова, индекс парадигмы, окончание исходной формы>.

Процедуре кластеризации были подвергнуты четыре массива:

а) основной тестовый массив НКРЯ;

б) 10000 наиболее частотных несловарных форм корпуса со снятой грамматической омонимией;

в) массив несловарных форм Яндекса (444 тыс.);

г) массив имен собственных НКРЯ (42 тыс.).

Для большого числа словоформ описанная в п. 2 процедура предсказала несколько гипотез с максимальным абсолютным весом. В случае, если предложенные варианты различались длиной основы, 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,

предпочтение отдавалось гипотезе с более длинной основой. У вариантов с совпадающей исходной формой и частью речи эта эвристика, в частности, позволила повысить вес гипотезы с более простым типом словоизменения относительно гипотезы с чередованием в основе (ср. Агриппин|а, S1_f:тип «тонна ~ тонн» и Агриппи|на, S15_f:тип «царевна ~ царевен»).

Прочие гипотезы с максимальным весом (различающиеся окончанием исходной формы и индексом парадигмы) признавались равновероятными, ср. гипотетические леммы «Мышкин», «Мышкина», «Мышкино» для словоформы «Мышкин»:

<Мышкин, Sf1_m, ∅>
<Мышкин, Sf1_f, а>
<Мышкин, S1_n, о>

При обработке массива имен собственных (г) порождались гипотезы о лемматизации словоформ только имен существительных и прилагательных.

4.4. Условности грамматической интерпретации лемм

Предложенный подход не позволяет определить одушевленность/неодушевленность леммы (для этого потребовался бы анализ синтаксического окружения словоформ в тексте). В связи с этим именам существительным приписывалась только грамматическая характеристика рода. Аналогичным образом, у глаголов не определялись переходность и вид. У имен собственных типа «Стенька», склоняющихся по женскому образцу, разбор «женский род» считается условно правильным.

Наречия и префиксоиды на -ски, -о (ср. «геополитически», «безапелляционно», «инновационно») попадают в кластеры лемм-прилагательных вследствие стандартной омонимии с краткими формами множественного числа и среднего рода. Несмотря на то, что вероятность интерпретации такой формы в реальном тексте как формы прилагательного близка к нулю, мы посчитали объединение форм прилагательных и производных от них наречий в общий кластер удобным для целей составления словника словаря. В случае необходимости такие формы могут быть выделены из кластера с помощью элементарной процедуры. То же самое касается форм сравнительной степени наречия на -ее, омонимичных форме среднего рода прилагательного, хотя тут ситуация обратная: в текстах по преимуществу представлены адъективные варианты.

В интерпретации словоформ на -ся, -сь практически всегда одинаковый вес имеют разборы возвратного и невозвратного глаголов. Поскольку формы, различающие парадигмы возвратного и невозвратного глагола, весьма редки (ср. «требуемся»), а дизамбигуация форм на -ся трудна даже для человека и обычно требует обращения к контексту, было принято техническое решение считать возвратную и невозвратную интерпретацию одним разбором (заметим, что подобный стандарт лемматизации принят в Частотном словаре Чешского национального корпуса [17]).

Понятие «условно правильной» кластеризации распространяется также на кластеры словоформ типа «Анджелес», «Анджелеса» и т. п. (вторая часть слова «Лос-Анджелес»), т. е. в том случае, когда ошибка лемматизации кроется в неправильных вводных данных.

5. Оценка эффективности алгоритма

5.1. Количественная оценка

Первый эксперимент по количественной оценке полученных данных предусматривал «ручную» проверку 1 тысячи наиболее частотных словоформ массива НКРЯ (а), вошедших в кластеры (табл. 2).

Таблица 2

1000 наиболее частотных кластеризованных разборов НКРЯ

Сл/форм на лемму	N кластеров	Разборов на сл/ф, в среднем	Ошиб. разборов:	
			сл/форм	кластеров
13	2	1		
12	4	1		
11	3	1	11	1
10	7	1		
9	10	1,1		
8	4	1		
7	9	1,22		
6	18	1,11	6	1
5	42	1,62		
4	57	1,91	4	1
3	28	6	3	1
2	4	2,5		

Кластеров с единственным правильным разбором: 92 (48,9 %);

– с правильным разбором и другими вариантами лемматизации: 91 (48,4 %);

– с ошибочным разбором (разборами): 5 (2,7 %).

Точность лемматизации словоформ составила 97,4 % при условии, что хотя бы один вариант лемматизации был правильным. Однозначно правильно было лемматизировано 59,2 % словоформ. В среднем на одну словоформу пришлось 1,46 разборов.

Эксперимент с лемматизацией массива НКРЯ (б) показал следующие результаты (табл. 3).

Таблица 3

10000 наиболее частотных разборов словоформ корпуса НКРЯ со снятой грамматической омонимией

Сл/форм на лемму	N кластеров	Разборов на сл/ф, в среднем	Ошиб. разборов:	
			сл/форм	кластеров
10	1	1		
9	2	1		
8	1	1		
7	3	1	2	1
6	4	1		
5	13	1,3		
4	39	1,46	6	3
3	114	6,55	75	25
2	702	5,6	205	106

Кластеров с единственным правильным разбором: 171 (19,5 %);

– с правильным разбором и другими вариантами лемматизации: 573 (65,2 %);

– с ошибочным разбором (разборами): 109 (15,4 %).

Покрытие массива кластерами: 20,5 %.

Точность лемматизации словоформ составила 85,9 % (хотя бы один вариант лемматизации правильный). Относительно всего корпуса однозначно правильно лемматизировано 496 словоформ, что составляет 24,2 % кластеризованных форм и 5 % исходного массива. В среднем на одну словоформу пришлось пять разборов.

Таким образом, второй эксперимент позволил обнаружить три слабых точки в исследуемом подходе: малое покрытие массива кластерами, неаккуратность лемматизации на трех- и двусловных кластерах, неоднозначность определения исходной формы и частеречных характеристик для словоформ, правильно объединенных в кластеры объемом от 2 до 5 элементов.

Мы предположили, что низкая степень покрытия массива кластерами объясняется тем, что был обработан не весь конкорданс слов по тексту, а лишь самая частотная его часть. Действительно, при увеличении объема данных в следующем эксперименте, с частотным словником Яндекса (в), кластеризация несловарных форм достигла 50 % (табл. 4).

Наилучшие результаты метод показал на массиве имен собственных НКРЯ (г), табл. 5.

Таблица 4

Кластеризация частотного словаря Яндекса, 444 тыс. несловарных словоформ

Словоформ на лемму	N кластеров	Нарастание покрытия, %
10 и более	1408	1,6
9	857	2,4
8	1306	3,5
7	2017	4,9
6	3532	7
5	7963	11
4	16062	17,4
3	34724	27,8
2	111003	50

Таблица 5

Кластеризация массива имен собственных НКРЯ, 42 тыс. несловарных форм

Сл/форм на лемму	N кластеров	Разборов на сл/ф, в среднем	Ошиб. разборов:	
			сл/форм	кластеров
14	2	1	3	2
13	2	1	4	2
12	4	1	3	2
11	8	1	15	3
10	12	1	6	3
9	23	1	10	5
8	41	1	6	3
7	53	1	11	5
6	147	1,09	32	22
5	598	1,31	113	73
4	1316	1,66	137	86
3	2015	2,69	328	180
2	5221	4,22	674	393

Кластеров с единственным правильным разбором: 2283 (24,2 %);

– с правильным разбором и другими вариантами лемматизации: 6196 (65,6 %);

– с ошибочным разбором (разборами): 963 (10,2 %) (усредненные данные: сравнивались результаты автоматической и ручной лемматизации для 2000 случайно выбранных кластеров).

Покрытие массива кластерами: 63,6 %.

Точность лемматизации словоформ 94 % (хотя бы один вариант лемматизации правильный).

5.2. Лингвистический анализ ошибок

Ошибки в лемматизации словоформ сводятся к двум типам: неправильное объединение словоформ в парадигму и неправильное определение исходной формы.

В принципе, исследуемый метод не учитывает возможной омонимии словоформ, присущей русскому языку (ср. форму «банка», входящую в пара-

дигму существительных «банк» и «банка»): предполагается, что результирующие кластеры не пересекаются друг с другом. Удачным образом, на периферии языка грамматическая омонимия практически отсутствует и системным образом проявляется лишь в зоне имен собственных.

В первую очередь, это пересечение парадигм притяжательных прилагательных, мужских и женских фамилий и топонимов среднего рода «Марфин/а/о», «Алабин/а/о», «Голицын/о», «Кунцев/о».

Прил.	Фамилия муж.	Фамилия жен.	Топоним (ср. р.)
Марфин Марфина ...			
Марфина Марфиной ...			
Марфино ...			
Марфины Марфиных ...			

На обработанном массиве данная особенность словоизменения проявилась в распознавании форм фамилий и топонимов как притяжательных прилагательных и в порождении «лишних» разборов на «-о». Поскольку корректная лемматизация словоформ в этом случае потребовала бы привлечения экстралингвистических знаний или анализа контекста, при оценке результатов такие варианты разбора признавались условно правильными.

Два других распространенных типа омонимии – пересечение парадигм мужских имен и фамилий по типу «Евграф/Евграфов» и пересечение парадигм прилагательных на «-ский/-цкий» и существительных на «-ск/-цк» – являются вырожденными, так как в текстах формы множественного числа имен маловероятны, а краткие формы прилагательных невозможны, хотя и не запрещены формально моделью словоизменения [16]. Случаи объединения в один кластер форм мужких и женских имен («Феодор/Феодора», «Раймонд/Раймонда», «Федосей/Федосья», по типу «статья»), а также имен и фамилий («Герда/Герд», «Фрида/Фрид») носят единичный характер.

Отдельную проблему представляют несклоняемые имена, которые «незаконно подключаются» к парадигме изменяемых существительных (ср. «Тито/Тит», «Коро/Корее», «Мозли/ Мозель», «Кольбе/Кольба», «Конго/Конг», «Конде/Кондей», «Тесье/Тесей», «Шани/(Тянь)-Шань», «Ронни/Ронен», «Корфу/Корф», «Клаудио/Клаудиа», «Густаво/Густав», «Левински/Левинский») или ошибочно объединяются с другими неизменяемыми формами (ср. «Пельш/ Пельше», «Анжи/Анжу», «Гете/Гетье», «Рене/Рено»). Ряд несклоняемых слов, представляющих «западно-ориентированные» варианты произношения, омонимичен формам склонения (ср. «Джульетт/Джульетта», «Эдит/Эдита», «Бонапарте/ Бонапарт», «Арктик/ Арктика»).

Значительное количество ошибок дает кластеризация словоформ по типу «римлян|ин» (ср. «Якоб/Якобина», «Франкл/Франклин», «Шульц/ Шульцину»). Представляется, что предварительная настройка матрицы русских окончаний под конкретную задачу, например, удаление окончаний краткой формы прилагательных или парадигмы «римлян/ин» при обработке массива имен собственных, могло бы повысить точность лемматизации.

Что касается нарицательных имен и глаголов, то ошибки объединения их словоформ в общую парадигму носят индивидуальный, несистемный характер и присутствуют обычно в двухсловных кластерах, ср. «слоган/сложа», «однакож/ однакоже», «разумем/разумну», «отче/отчину», «быти/бытье», «невыданные/невыдающаяся». Значительно выше риск некорректной кластеризации в парах имя нарицательное – имя собственное и имя нарицательное – аббревиатура, ср. «було/Буденный», «кантри/Кантер», «маман/ Мамай», «кунать/Куня», «грит/ГРУ», «тово/тов.». Сравнение результатов обработки массива несловарных слов Яндекса и массива нарицательной лексики НКРЯ показало, что выделение имен собственных и аббревиатур в отдельные множества позволяет избежать таких псевдо-парадигм.

Ошибки, связанные с неправильным определением исходной формы, можно проиллюстрировать примерами «остывнуть, V» (формы «остыв», «остывшая», «остывшего» и т.д), «накопительной, S_m» (формы «накопительное», «накопительной»), «Громыкий, A» (формы «Громыко», «Громыки»). Как правило, аккуратное определение границ кластера гарантирует его корректную лемматизацию (один разбор из предложенных с максимальным весом является правильным). Однако изредка возникает побочный эффект от применения эвристики «длинной основы»; кроме того, программа некорректно обрабатывает имена с нестандартным типом склонения.

В целом, эвристика «длинной основы» показала хорошую эффективность на обработанном материале, понижая вес разборов с чередованием и вес глагольных разборов. В частности, следствием ее применения стала лемматизация отглагольных образований типа «неработающий», «неоконченный» и «полусгнивший» как прилагательных, а не как глаголов типа «неработать», «полусгнить». Вместе с тем, как прилагательные были разобраны нестандартные формы причастий типа «привыкш|ий/привыкш|его» (ср. «привыкнувший»), а также отдельные формы глаголов, которые встретились в массиве только в причастной форме («аффилированный»). У некоторых прилагательных была неправильно проведена граница между основой и окончанием (вариант «инновационн|ый» получал больший вес, нежели «инновацион|ный»)[3]. Недостаток эвристики проявлялся в тех случаях, когда в кластере объединялись словоформы с беглой гласной в основе типа «палестинцы», «палестинцев» (лемма «палестинц» вместо «палестинец»). Наконец, в двусловных кластерах встретились такие разборы, как существительные

мужского рода «запостил» (формы «запостил», «запостила»), «гранулометрическом» (формы «гранулометрическом», «гранулометрическому»), «накопительной» (формы «накопительной», «накопительное»).

Ошибки в определении исходной формы были связаны также с неполнотой исходных грамматических данных. В матрице русских окончаний не были учтены парадигмы pluralia tantum («Борки», «Помпеи», «Кукрыниксы») и некоторые нестандартные парадигмы типа «сверхновая» (субстантивированное прилагательное), «Василенко/Василенки», «Василий/Василья/ Василью», «Клаудиа/Клаудии» (ср. также «Икеа», «Нивеа»), «божий/божиих», нестандартные окончания «-ьи» («в молчаньи»), «-ию» («добродетелию»), «-ьми» («добродетельми»). Впрочем, доля таких экзотических кластеров в общем массиве была крайне мала, а значит, увеличение числа парадигм никак не сказалось бы на эффективности работы программы.

5.3. Проблема неоднозначности лемматизации

Ранее мы критиковали словарно-ориентированный метод лемматизации за то, что неопознанным формам приписывается в среднем три разбора. Как видно, метод парадигматической лемматизации также не решает проблему снятия грамматической омонимии: на одно несловарное слово здесь может приходиться до шести и более равноправных гипотез. Причиной такого положения является сильное пересечение множеств окончаний у сущест-вительных трех родов, прилагательных на «-ый»/«-ий» и «-ой» и глаголов.

По нашим эмпирическим оценкам, для дизамбигуации разборов разных частей речи можно было бы воспользоваться следующей иерархией приоритета:

существительное > прилагательное > глагол.

Чаще всего между собой конкурировали леммы одной части речи, а именно, две-три гипотезы о существительном мужского, среднего и женского рода или две гипотезы о прилагательном с окончанием «-ый»/«-ий» ~ «-ой». С точки зрения теории, соответствующие парадигмы различаются «сигнальными» формами творительного падежа единственного числа на «-ой» (ж. р.), «-ом» или «-ым» (м. и с. р.), родительного падежа множественного числа на «-ов» (м. р.) и т. д., между тем как на практике в кластеры попадают самые «распространенные» формы с окончанием, например, «-∅», «-а», «-е». Таким образом, проблема неоднозначности в целом остается неразрешенной и снижает точность обработки результатов. Заметим, впрочем, что существует определенная корреляция между неоднозначностью лемматизации и объемом кластера: вышеописанная омонимия наиболее характерна для двух- и трехсловных кластеров.

[3] Эта ошибка не является критической (неверно предсказывается краткая форма мужского рода «инновационн» «инновационен»), поэтому такой разбор признавался условно правильным.

6. Выводы и перспективы

В данном проекте метод парадигматической лемматизации использовался для составления словаря новых слов (лексем изменяемых частей речи) на базе массива несловарных словоформ. Алгоритм был испытан на четырех массивах несловарных слов разного объема и состава.

В ходе работы программы массив словоформ разбивался на кластеры, представляющие парадигмы слов изменяемых частей речи, и остаток, представленный неизменяемыми словами и «изолированными» словоформами. Метрика полноты измерялась как степень покрытия массива кластерами: этот показатель варьировал от 20,5 до 63,6 %. Степень покрытия зависела от объема исходных данных: низкий результат был получен на массиве 10 тыс. словоформ, удовлетворительные результаты – на массивах 42 тыс. и 444 тыс. словоформ.

Точность лемматизации составила от 85,9 до 97,4 % при условии, что хотя бы один вариант лемматизации признавался правильным. Основной проблемой метода признана неоднозначность лемматизации, которая ярче всего проявляется в кластерах объемом от двух до трех словоформ. На кластерах объемом четыре словоформы и более алгоритм стабильно выдавал правильные и по преимуществу однозначные результаты.

Была показана также зависимость эффективности метода от качества исходного массива: предварительное выделение имен собственных и аббревиатур, основанное на информации о капитализации и пунктуационном окружении словоформ в тексте, позволила более тонко настроить процедуру лемматизации и добиться оптимальных результатов.

В настоящее время стали известны результаты еще трех экспериментов по лемматизации несловарных слов, полученных в дипломной работе Д. К. Бронниковой [19] под руководством авторов проекта. В частности, исследовался вопрос, можно ли добиться улучшения статистики путем объединения усилий словарно-ориентированных анализаторов (Mystem, Starling) и программы парадигматической лемматизации. Ответ на этот вопрос скорее неутешительный: так, на массиве из 388 словоформ, имеющих несколько вариантов лемматизации по версии Mystem, алгоритм парадигматической лемматизации выбрал единственный правильный разбор только для 191 словоформы (49,2 %). Более того, оказалось, парадигматический метод работает тем успешнее, чем больше гипотез для каждой словоформы порождается; анализаторы Mystem и Starling, напротив, ограничивали выбор гипотез и тем самым снижали эффективность работы парадигматического алгоритма.

Более эффективный способ улучшения лемматизации заключался в отбрасывании продуктивных префиксоидов типа «авиа-», «авто-» и проверке вхождения «хвоста» в существующий словарь. Был составлен словарь из 89 словообразующих корней, и с его помощью удалось лемматизировать 73 % сложных слов с точностью более 97 %.

Эксперимент с префиксоидами доказывает, что методы, построенные на лингвистических правилах, превосходят статистические методы по параметру точности (но не полноты). В ходе исследований мы предложили ряд других лингвистических эвристик, дополняющих метод парадигматической лемматизации и способствующих повышению его точности: подавление гипотез о краткой форме прилагательного, о редких типах склонения с чередованием, иерархия приоритетов и др. Их эффективность предстоит проверить в будущих экспериментах.

В заключение хотелось бы сказать, что кажущийся небольшим коэффициент кластеризации – до 63,6 % массива – на самом деле обозначает не так уж мало. На массиве в 1 млн словоформ данный метод позволяет установить порядка 180 тыс. потенциальных лемм, а этот показатель превышает объем большинства электронных грамматических словарей.

7. Благодарности

Авторы выражают благодарность Е. В. Рахилиной, В. А. Плунгяну, принимавшим участие в обсуждении идеологии проекта, Е. А. Гришиной, оказавшей неоценимую помощь в разрешении трудных вопросов лемматизации, и в особенности Г. К. Бронникову и Д. К. Бронниковой, участвовавшим в дискуссии относительно методов реализации алгоритма и предоставившим нам материалы своих исследований.

8. Литература

[1] *Mikheev A.* Automatic rule induction for unknown word guessing // Computational Linguistics. 1997. Vol. 23. № 3. P. 405–423.

[2] *Сегалович И., Маслов М.* Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // Тр. междунар. семинара Диалог'98 по компьютерной лингвистике и ее приложениям. – Казань, 1998. Т. 2. – С. 547–552.

[3] *Сокирко А. В.* Морфологические модули на сайте www.aot.ru // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2004». – М., 2004.

[4] *Trost H., Buchberger E.* Towards the Automatic Acquisition of Lexical Data // The 11th International Conference on Computational Linguistics, Itonn, Germany, 1986. P. 387–389.

[5] *Wilensky R.* Extending the Lexicon by Exploiling Subregnlarities // The DARPA Speech and Natural Language Workshop, Llidden Valley, Pennsylvania, 1990. P. 365–370.

[6] *Daciuk J.* Computer-assisted enlargement of morphological dictionaries: Finite state methods in natural language processing // Workshop at 13th ESSLLI. Helsinki, 2001.

[7] *Viegas E., Onyshkevych B. A., Raskin V., Nirenburg S.* From Submit to Submitted via Submission: On Lexical Rules in Large-Scale Lexicon Acquisition // ACL 1996. P. 32–39.

[8] *Коваленко А.* Стемка – морфологический анализ для небольших поисковых систем // Системный администратор. 2002. № 1, окт.

[9] *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.

[10] *Hana J., Feldman A.* Portable language technology: Russian via Czech // Proceedings of the Midwest Computational Linguistics Colloquium, 2004, Bloomington, Indiana.

[11] *Сокирко А. В.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка. [Электрон. ресурс]. 2005. Режим доступа: http://company.yandex.ru/grant/2005/01_Sokirko_92802.pdf

[12] *Зеленков Ю. Г., Сегалович И. В., Титов В. А.* Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конф. Диалог'2005. М., 2005. http://www.dialog-21.ru/Archive/2005/Zelenkov%20Segalovich/Zelenkov_Segalovich.pdf.

[13] *Ножов И. М.* Реализация автоматической синтаксической сегментации русского предложения. Дис. ... канд. тех. наук. – М.: РГГУ, 2003.

[14] *Kanis J., Müller L.* Automatic Lemmatizer Construction with Focus on OOV Words Lemmatization // Text, Speech and Dialogue 2005. Berlin/Heidelberg: Springer, 2005. P. 132–139.

[15] *Бидер И. Г., Большаков И. А., Еськова Н. А.* Формальная модель русской морфологии. Ч. 1–2. М., 1978.

[16] *Зализняк А. А.* Грамматический словарь русского языка: Словоизменение. – М., 1977; 4-е изд. – М., 2003.

[17] *Ляшевская О. Н.* К проблеме лемматизации несловарных слов // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конф. Диалог'2007. М., 2007.

[18] *Čermák F., Křen M.* (eds.). Frekvenční slovník češtiny – Praha: NLN, 2004.

[19] *Бронникова Д. К.* Сравнение алгоритмов лемматизации на материале Национального корпуса русского языка. Дипломная работа. – М.: РГГУ, 2007.

Automatic enlargement of a dictionary: from the set of unknown word forms to the lemmatized list

O. Lashevskaja, D. Sichinava, B. Kobritsov

Text tokens that are not represented in the dictionary of a morphological parser pose a problem both for the automatic analysis of texts and for the compiling of corpora-based dictionaries. We evaluate an algorithm according to which unknown word forms are grouped in clusters and associated with part of speech, base form and other grammatical information. The clusterization procedure consists in generation of multiple hypotheses for each word form in compliance with A.A. Zalizniak's Russian derivational model and weighting up the frequency of hypotheses throughout the whole domain.

The evaluation of the algorithm efficiency is set up on the concordance of the Russian National Corpus and the dataset 'Yandex bank of word forms'.