Быстрословарь: предсказание морфологии русских слов с использованием больших лингвистических ресурсов

Bystroslovar': morphological prediction of new Russian words using very large corpora

Сокирко А. В. (sokirko@yandex-team.ru)

ООО «Яндекс», Москва, Россия

Рассматривается система предсказания морфологических признаков и полной парадигмы слова, разработанная для поиска по веб-страницам на русском языке. Система использует 12 факторов предсказания, размеченный корпус примеров, на котором тренируется модель машинного обучения. Описаны факторы, на которых строится обучение, делается попытка оценить значимость каждого фактора для решения поставленной задачи. Приводится сравнение построенной модели по точности и полноте с моделью, построенной на одном факторе.

1. Постановка задачи

Роль крупных Интернет-порталов, предоставляющих пользователям возможность поиска по вебу, растет с каждым годом. Ранжирование веб-страниц по релевантности пользовательскому запросу не решается без установления того, что слова из пользовательского запроса равны по значению словам, взятым с веб-страницы. Самое частое преобразование слов при поиске в рунете — морфологическое. Традиционно большая часть морфологической информации записывается в словаре, это удобно, поскольку морфология языка достаточно статична. Для каждого входного слова в этом словаре даются морфологические характеристики и парадигма словоизменения. Формально для поиска нужна только парадигма словоизменения, но в реальности определение, например, части речи может помочь установлению всех форм слова. Кроме статичного основного словаря, во многих системах строится специальный модуль предсказания новых слов, которых нет в основном словаре. Поскольку веб-поиск в каждый момент времени имеет дело с фиксированным («замороженным») корпусом, представляется целесообразным спроектировать словарь, функционально равный основному морфологическому словарю, но построенный автоматически по пользовательским запросам и веб-страницам. Такой словарь мы будем называть «быстрословарем», поскольку строится быстрее ручного словаря и поскольку он должен обновляться достаточно часто. В идеале этот словарь должен включать все слова, найденные в рунете и не вошедшие в основной словарь.

Данная публикация описывает новый алгоритм построения быстрословаря. Основным достоинством этой версии мы считаем:

1. Учет многих факторов морфологического предсказания внутри одной системы машинного обучения, что для русского языка не является разработанным приемом.
2. Получение информации из агрегированных пользовательских запросов и из текстов веб-страниц.

2. Обзор литературы

Системы морфологического анализа — одна из самых разработанных областей компьютерной лингвистики. Морфологические системы для русского языка традиционно основываются на грамматическом словаре А. А. Зализняка[1]. Алгоритмы предсказания обобщают морфологические схемы, предложенные в этом или родственном ему словарях, и/или учитывают распространенность этих схем в словаре и в обучающем корпусе. Один из первых алгоритмов морфологического предсказания для русского языка был предложен в работах Г. Г. Белоногова [2, 3]. Главным в этом алгоритме предсказания был принцип «корреляции между грамматическими признаками слов и буквенным составом их кон-

цов». Дальнейшее развитие алгоритмов, описанное, в частности, в работах Гельбуха[4], включало более детальную проработку правил композиции морфем и учет статистики морфем в корпусе текстов.

В 80-е и 90-е годы на факультете ВМК МГУ активно разрабатывалась система TULIPS-2, которая включала морфологический компонент[5], эта система использовала для предсказания словарь основ и словарь флексий, учитывались чередования.

Развитие корпусной лингвистики подстегнуло рост интереса к системам, которые в качестве решающего фактора используют частотность тех или иных морфологических схем в текстовом корпусе. Например, в работе Wicentowski[8] исследуется система, которая построена на трех простых факторах:

1. Расстояние Левенштейна, модифицированное под поиск морфологических вариантов
2. Контекстная близость по соседним словам в корпусе
3. Близость по частоте форм в одинаковых моделях словоизменения. Показывается, что система дает точность лемматизации порядка 80 % на 30 различных языках.

В работе Ножова [9] предлагается взвешивание гипотез морфологического предсказания с использованием метод корреляции. Матрицы корреляции строятся для основ и значений классифицирующих грамматических категорий. Гипотезы, имеющие максимальную корреляцию объявляются наиболее правдоподобными.

В работе Ляшевской и др.[10] был предложен метод взвешивания морфологического предсказания, основанный на следующем утверждении.

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. Из этого можно сделать предположение, что парадигмы новых слов тем лучше, чем больше разных форм этой парадигмы найдено в корпусе. В этой работе строились парадигмы для слов из Национального корпуса русского языка (НКРЯ).

В последних работах Goldsmith[11], одного из самых увлеченных исследователей в этой области, исследуется принцип минимального описания, который гласит, что морфологическая теория, которая описывает тот или иной корпус, должна быть минимальной по длине. Например, если в корпусе можно выделить N основ и K флексий, тогда такая теория лучше, чем теория, которая выделяет N+1 основ или K+1 флексий.

Хочется отметить, что многие работы в этой области затрагивают два аспекта, которые нас совершенно не интересуют:

1. Теоретическая морфология, когда авторы пытаются смоделировать очень глубокие законы, которые на сегодняшний день уже не являются продуктивными или распространенными. Мы считаем, если есть возможность задать «исключения» (следствие «устаревших» законов) списком, это достаточно.
2. Проблемы алгоритмической эффективности определения новых слов (нас интересует исключительно качество найденного, но не скорость работы программы).

3. Общая схема предсказания

Итак, главная задача — построить автоматический словарь («быстрословарь») по образу основного словаря. Основной словарь — это словарь Mystem[12]. Для нашей задачи можно считать, что это просто модифицированная версия грамматического словаря Зализняка. Каждое слово описывается набором форм и морфологическими характеристиками, которые приписаны этой форме. Например, для слова «мама»:

мама S,од,жен,ед, им,
мамы S, од, жен,ед, род
маме S,од,жен,ед, дат
маму S, од,жен,ед,вин
мамой S, од, жен, ед,твор
маме S,од,жен,ед,пр
мамы S,од,жен, мн,им
....

В каждой такой парадигме можно выделить псевдооснову (неизменяемую левую часть), в данном случае *мам-*, можно выделить StemGrammar (словообразовательные пометы, в данном случае «S,од,жен») и FlexGrammar (словоизменительные пометы). Можно записать данную парадигму в виде тройки <Основа, StemGrammar, Модель окончаний>, где модель окончаний — это набор пар вида <окончание, FlexGrammar>, например:

мама = <*мам*, «S,од, жен», F>,
где F = <-*а*, ед,им>, <-*ы* ед,род>..

В текущей версии словаря используются около 3000 моделей. Некоторые из них уникальны, например, есть специальная модель для слова *Комсомольск-на-амуре*, там выделяются окончания *-а-на-амуре*, *-ом-на-амуре* и т. д.

Предполагается, что в быстрословаре не будет отличных от основного словаря наборов окончаний или StemGrammar, предполагается даже, что самые редкие модели окончаний или StemGrammar будут убраны из рассмотрения.

Кроме основного словаря, используется еще два источника. Это веб-страницы рунета (около 4 миллиардов русских страниц без спама, только html) и лог частотных пользовательских запросов, где для

каждого запроса указана частота за месяц. Топ этого списка (всего 400 млн. запросов) выглядит так:

одноклассники	3 308 624
в контакте	1 457 124
порно	1 129 705
mail.ru	1 063 249
вконтакте	690 068
контакт	558 708
погода	458 272
зайцев нет	441 107
однокласники	391 875
работа	356 324
vkontakte	350 099
из рук	*в руки* 325 295
гороскоп	309 147
википедия	307 352
рамблер	303 245

По вебу и по запросам составлены частотные словари, где указана уже статистика для отдельных слов.

Кроме входных данных, есть еще корпус размеченных запросов (т. н. «морфотест»), в котором для каждого слова запроса указана вся парадигма слова. Корпус содержит 10 тысяч запросов и размечен вручную.

Общая схема построения словаря выглядит так. Мы берем все словарные слова из морфотеста, кроме стоп-слов и самых частых слова. Для каждого слова строятся все возможные морфологические интерпретации, хотя известно, что только одна из них верная. Для всех интерпретаций строятся значения всех факторов предсказания, и все вместе это составляет обучающую выборку.

Например, для слова *мама* строятся варианты (гипотезы):

Example	ModelId	StemGrammar	FlexLen	ModelFreq	IsGood
рома	7	S,имя,муж,од	1	212	–
Дюма	1135	S,од,фам	0	234	–
Фатима	7	S,жен,имя,од	1	2754	–
Айова	23	S,муж,неод,гео	1	13 908	–
ведьма	7	S,жен,од	1	1000	+
…					

Столбец Example — это пример, по которому была построена данная гипотеза. ModelId — номер модели окончаний (по нему строятся все формы слова). FlexLen — длина псевдоокончания, ModelFreq — частота модели окончаний с данным StemGrammar в основном словаре. Столбцы FlexLen и ModelFreq даны как примеры факторов предсказания, по которым происходит обучение (подробнее ниже). Столбец IsGood содержит целевой фактор обучения, он равен ‘+’ только для верной гипотезы («S, жен, неод», ModelId=7). На самом деле в реальной модели целевой фактор является не булевским, а некой шкалой, поскольку нас интересует не точное попадание в единственно верную модель окончаний, а хотя бы приблизительное. В обучающую выборку было взято 4972 слова, по ним было построено 124 677 гипотез. Тестовая выборка состояла уже из несловарных слов морфотеста (998 слов), по ним было построено 24 923 гипотезы. Здесь может возникнуть вопрос: почему обучающая выборка состоит из словарных слов, а тестовая из несловарных? Очевидно, что выборки различны (см. об этом более подробно в работе Клышинского[13]). К сожалению, нам приходится мириться с этим, поскольку выборка по несловарным словам слишком мала для выделения из нее обучающей части.

Дальше на обучающей выборке строится модель машинного обучения, которая может предсказать StemGrammar и ModelId. На тестовой выборке проверяется качество модели. После этого мы берем топ самых частых несловарных слов из запросов (около 1,5 млн), строим для них все гипотезы, модель машинного обучения предсказывает StemGrammar и ModelId, по самым лучшим предсказаниям строятся парадигмы слов, если парадигмы сильно пересекаются друг с другом или со словарными словами, происходит отсев худших гипотез. Полученное множество парадигм и есть новый быстрословарь.

4. Факторы предсказания

Отдельный фактор должен голосовать за ту или иную гипотезу морфологического предсказания. Гипотеза морфологического предсказания (как сказано выше) — это модель окончаний (ModelId) и словообразующие характеристики (StemGrammar). Нет необходимости понимать, как отдельный фактор участвует в формуле предсказания, достаточно знания, что значение этого фактора **влияет** на выбор морфологической интерпретации. Оценка влияния фактора (importance, важность) возникает позже, после автоматического подбора формулы. Конечно, оценка Importance может только приблизительно показать значимость фактора, и она очень сильно зависит от выбранной модели машинного обучения. В нашем случае Importance дается по модели RandomForest[14], построенной на 30 решающих деревьях (реализация пакета R[15]).

5. Факторы внутреннего состава слова

Исторически самые используемые факторы для русского языка были связаны с окончанием сло-

ва (=псевдоокончание или постфикс). В нашей работы рассматриваются постфиксы длиной от 1 до 6 символов. По входному слову и гипотезе предсказания строится, соответственно, не больше 6 постфиксов, каждый из которых оценивается с помощью следующих параметров: Fp(p) — сколько раз постфикс p был использован для данной модели окончаний во всем словаре, Fg(p) — сколько раз постфикс p был использован с данным StemGrammar во всем словаре, Fa(p) — сколько раз постфикс p был использован во всех словах словаря. С помощью этих частот строятся следующие факторы:

Suffix_g — сумма всех Fg(p)/Fa(p) для всех постфиксов слов предсказываемой парадигмы.
BadSuffix_g — количество всех постфиксов, для которых Fg(p)=0 для всей предсказываемой парадигмы.

Suffix_p, **BadSuffix_p** — аналогично Suffix_g и BadSuffix_p, но вместо функции Fg используется Fp. Получается, что **Suffix_g** и **BadSuffix_g** оценивают, насколько слово подходит предсказываемому StemGrammar, а **Suffix_p**, **BadSuffix_p —** предсказываемой модели окончаний.

Важность этих факторов для предсказания StemGrammar высока:

Фактор	Предсказание StemGrammar	Предсказание ModelId
Suffix_g	20,21 %	7,09 %
Suffix_p	2,54 %	6,40 %
BadSuffix_g	8,07 %	2,84 %
BadSuffix_p	3,12 %	3,11 %

Суммарная важность этих факторов равна 33 % для предсказания StemGrammar и 20 % для предсказания ModelId.

Кроме этого, еще использовались другие факторы, описывающие внутренний состав слова:

- **FlexLen** — длина псевдоокончания для предсказываемой формы внутри предсказываемой парадигмы.
- **Hasprefix** — в словаре найдено другое слово W той же модели окончания, такое что входное слово делится на две части P и W, где P — продуктивный префикс типа «квази», «экс» и т. д.
- **lemma_leven** — в словаре найдено другое слово W той же модели окончания такое, что расстояние Левенштейна между входным словом и словом W не больше некоторого порога.
- **AbbrLike** — отношение числа согласных букв слова к общему количеству букв слова.

Каждый из этих пяти факторов имеет важность не больше 2 %. На данном этапе не очень понятно, почему их вклад такой незначительный.

6. Контекстные факторы

Следующие два фактора используют данные про контексты слов, взятые с web-страниц. Под обработку контекстов была специально адаптирована модель Trigram[16], построенная на полных тегах и обученная на Национальном корпусе русского языка. Для каждого контекста входного несловарного слова строятся вероятности того, что этому слову может быть приписана данная морфологическая интерпретация (StemGrammar и FlexGrammar). При использовании этих факторов возможны следующие осложнения. Во-первых, тексты НКРЯ не равны текстам рунета по жанрам и среднему количеству ошибок, хотя это не так критично, ведь модель Trigram основывается на тройках морфологических интерпретаций, типа

«A,ед,муж,им»,
«A,ед,муж,им»,
«S,ед,муж,им» *// красивый красный шар*
// новый желтый стол

Такие триграммы широко распространены как в НКРЯ, так и в рунете.

Во-вторых, некоторые контексты в рунете столь неоднозначны, что опираться на них не стоит. Учитывая это, мы пропускаем те контексты, в которых уровень морфологической неоднозначности превышает некий порог. В-третьих, не все тексты одинаково хороши, некоторые сайты содержат много мусора. Оценить значимость сайта можно с помощью т. н. тематического индекса цитирования[17], но пока этого не сделано.

В текущую модель включено два контекстных фактора:

- **weight_context_exact** — количество контекстов рунета, которые проголосовали за морфологическую интерпретацию входного слова;
- **weight_context_sum** — считаем **weight_context_exact** для каждой формы парадигмы и нормируем на число форм в парадигме.

Суммарная важность этих двух факторов находится в районе 12 %.

7. Факторы частот

Следующая группа факторов связана с частотными словарями рунета и пользовательских запросов:

- **QF** — частота входного слова в пользовательских запросах;
- **TF** — частота входного слова в рунете;
- **PROP** — отношение числа раз, когда слово было записано в рунете с большой буквы, к значению TF;

- **QF1** — частота слова в однословных запросах;
- **PrdFreqModel** — среднее квадратичное отклонение относительных частот форм этой модели окончания. Например, в парадигме слова «мама» (ModelId=7), относительная частота словоформы *мама* — 30 % , а словоформы *маме* — 10 % и т. д. Если в этой модели (ModelId=7) средние относительные частоты совсем другие — тогда штрафуем эту гипотезу.
- **paradigm_found_pure** — число форм парадигмы, найденных в рунете, поделенное на число форм парадигмы.
- **lemma_query_freq** — число вхождений предсказываемой леммы парадигмы в лог пользовательских запросов.

Суммарная важность этих факторов для StemGrammar — 25 %, а для ModelId — 35 %. Именно по этим группам факторов происходит основное расхождение между словарными и несловарными словами, поскольку, например, словарные в среднем используются в рунете на порядок чаще, чем несловарные.

8. Остальные факторы

Поскольку среди морфологических помет есть несколько, которые не относятся непосредственно к морфологическим (гео, имя, фам), нам пришлось подключить два источника:

1. Тезаурус географических названий Яндекс.Карт.
2. Частотные словари сервиса МойКруг для фамилий и имен.

На основе этих перечней мы построили фактор **SemFeat**, который, например, поднимает гипотезы с пометой «фам», если это слово часто встречается на сервисе МойКруг. Важность этого фактора не оказалась высокой даже для предсказания StemGrammar (1 %).

Кроме этого, мы использовали еще два дополнительных технических фактора:

- **ModelFreq** — частота модели окончаний в основном словаре (важность — 2 %).
- **FormsCount** — число форм в предсказываемой парадигме (важность — 20 %).

9. Анализ результатов

Построение всех факторов занимает порядка 20 часов на кластере в 200 компьютеров (каждый компьютер — 8 процессоров, 20 ГБ ОЗУ, 5 ТБ диск). Основное время уходит на подсчет контекстных факторов.

Сравнение по качеству производилось по двум метрикам:

1. Точность определения StemGrammar;
2. Точность и полнота определения форм парадигмы.

Последние определялись как средние точность и полнота для каждого слова запроса:

$$Точность: P = C/V,$$

$$Полнота: R = C/Q,$$

где C — сумма частот(в рунете) правильно предсказанных форм парадигмы, V — сумма частот всех предсказанных форм парадигмы, Q — общая сумма частот всех правильных форм парадигмы.

$$F\text{-}Measure: Fm = 2*P*R/ (P+ R)$$

Для сравнения была использована предыдущая версия быстрословаря, которая по большей части базировалась на работе[10]. Ниже приведены показатели качества:

	Точность	Полнота	F-measure
StemGrammar, old	0,6716		
StemGrammar, new	0,7517		
Формы, old	0,9295	0,9372	0,9354
Формы, new	0,9282	0,96	0,9439

Прирост точности по StemGrammar достаточно большой (с 67 % до 75 %), это вполне соответствует нашим ожиданиям, ведь мы перешли от довольно простой модели (которая включена в нашу модель в виде одного из фактора частот) к многофакторной модели, учитывающей многие аспекты использования слова.

Показатели по формам не столь оптимистичны. Общая F-measure выросла на процент, однако точность немного снизилась (это не мешает нам говорить о том, что модель в целом лучше предыдущей). Почему точность снизилась? Ниже приведен список самых грубых ошибок нового быстрословаря на текущей версии морфотеста:

Дробо	Предсказано как краткое прилагательное, должно быть фамилией
пожаро	Должно быть в основном словаре
стил	Должно быть неизм. фам, а стало изм., мешают формы *стиле*
руссо	Стало изменяемым, формы мешаются с фамилией *Русс*
Мега	Иногда изм, иногда нет
Дони	предсказалось как имя *Доня*, а в запросе это часть детской считалки
Волошко	не стало фамилией

Куинджи	Предсказано как изменяемая фамилия
Макроуровне	предсказалось как отчество по *Артуровне*
Рогово	Предсказалось как прилагательное
Судоку	Предсказалось как изменяемое

Даже этот короткий список примеров может определить направления будущей работы:

1. Проблемы с «Пожаро». Основной словарь требует доработки, не расширения, а упорядочивания. Например, мы последовательно должны включать композитные формы в парадигмы прилагательных.
2. Многие формы в интернете (судоку, куинджи) становятся изменяемыми. Например, таковым стало слово «вконтакте» (вконтакту, «вконтактом»…). Провести границу между совсем уже редким или полушуточным использованием или привычным освоением этого слова трудно, но нужно.
3. Пересечение слов со словарными и предсказанными парадигмами (*стил*, *руссо*), особенно для коротких слов, требует отдельной проработки. Фактически почти все предложенные факторы — это факторы для жадного захвата форм в парадигмы, но должны существовать факторы, которые препятствуют объединению несовместимых форм. Единственный фактор, который используется для этого, — это PrdFreqModel. Возможно, нужно вводить еще факторы, например, сколько раз разные формы слов одной парадигмы встречались на одной веб-странице.
4. По всей видимости, любые контекстные факторы требуют домножения на оценку качества самой веб-страницы.

Кроме тактических соображений, еще предстоит начать переоценку масштабов заимствования и скорости освоения новых слов рунета.

Благодарности

Я выражаю благодарность всем сотрудникам отдела лингвистических технологий компании «Яндекс» за помощь в проведении этого исследования, а особенно:

1. Евгению Соловьеву (машинное обучение);
2. Николаю и Светлане Григорьевым (разработка системы оценки качества);
3. Елене Грунтовой (общее руководство и идея исследования);
4. Вере Цукановой (основной морфологический словарь);
5. Андрею Кондратьеву (предыдущая версия быстрословаря).

Литература

1. *Зализняк А. А.* Грамматический словарь русского языка. Словоизменение. — М.: Русский язык, 1977.
2. *Белоногов Г. Г.* Об использовании метода аналогии при автоматической обработке текстовой информации // Проблемы кибернетики. — М.: 1974, вып. 28.
3. *Белоногов Г. Г., Зеленков Ю. Г.* Алгоритм морфологического анализа русских слов // Вопросы информационной теории и практики. № 53. Автоматическая словарная служба. Автоматическое индексирование документов. М., 1985. С. 62–93.
4. *Гельбух А. Ф.* Эффективно реализуемая модель морфологии флективного естественного языка: Автореф.дис… к. ф. н. / АНРФ., ВИНИТИ. — М.,1994.
5. *Мальковский М. Г., Волкова И. А.* Анализатор системы TULIPS-2. Морфологический уровень // Вестн. Моск. Ун-та, сер. 15, 1981, N 1, с. 70–76.
6. *Шереметьева С. О., Ниренбург С.* 1996 Эмпирическое моделирование вычислительной морфологии. //НТИ, №7, 1996.
7. *Goldsmith, J.* Unsupervised Learning of the Morphology of a Natural Language. //University of Chicago, 1998.
8. *Wicentowski, R.* Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework, 2002.
9. *Ножов И. М.* Реализация автоматической синтаксической сегментации русского предложения. Дисс… канд. тех. наук. — М.: РГГУ, 2003.
10. *Ляшевская О. Н., Сичинава Д. В., Кобрицов Б. П.* Автоматизация построения словаря на материале массива несловарных словоформ // Браславский П. И. (отв. ред.), Интернет-математика — 2007: сб. работ участников конкурса науч. проектов по информ. поиску. Екатеринбург: Изд-во Урал. ун-та. С. 118–125.
11. *Goldsmith, J.* 2001. The unsupervised learning of natural language morphology. Computational Linguistics 27(2): 153–198.
12. *Сегалович И., Маслов М.* Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // Труды международной семинара Диалог'98 по компьютерной лингвистике и ее приложениям. Казань, 1998. Т. 2. С. 547–552.
13. *Клышинский Э. С.* Некоторые сложности автоматизированной лемматизации несловарных словоформ [Текст] [Электронный ресурс] / Э. С. Клышинский // Материалы международной конференции «Диалог 2009». — М.: РГГУ, 2009
14. *Breiman, L.* Random forests. Machine Learning, 45(1): 5–32, 2001. 18
15. *Liaw, A., Wiener, M.* Classification and regression by randomForest. Rnews 2002, 2:18–22.
16. *Сокирко А. В., Толдова С. Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка // Интернет-математика-2005.
17. *Что такое тИЦ?* Электронная публикация. http://help.yandex.ru/catalogue/?id=873431
18. *Toutanova, K., Cherry, C.* 2009:A global model for joint lemmatization and part-of-speech prediction