УДК 81 + 81'32 + 81'33
DOI: 10.17223/19986645/68/2

Е.В. Люкина, М.А. Лытаева

ИНКРЕМЕНТАЛЬНОЕ ПРЕДСКАЗАНИЕ МОРФОЛОГИЧЕСКОЙ ПАРАДИГМЫ НЕИЗВЕСТНЫХ СЛОВ РУССКОГО ЯЗЫКА

Рассматривается метод предсказания морфологической парадигмы неизвестных (несловарных) слов русского языка, позволяющий в инкрементальном режиме, т.е. по мере поступления очередной словоформы, в полностью автоматическом режиме, без участия человека, предсказать морфологическую парадигму слова. Исследование проводилось на корпусе текстов русского языка различной тематики (более 10 млрд словоупотреблений), позволяющем изучить эффективность предлагаемого способа предсказания морфологической парадигмы слов.

Ключевые слова: морфология, морфологический анализ, морфологическая парадигма, предсказание морфологической парадигмы неизвестных слов, вычислительная лингвистика, корпусные исследования.

Введение

Во многих задачах анализа текстов естественного языка, таких как синтаксический разбор текста, классификация текста, корпусные исследования и пр., одним из важнейших этапов является морфологический анализ.

Морфологический анализ предназначен для получения морфологических признаков слова. Приведем некоторые термины морфологического анализа. Текст на естественном языке задается в виде упорядоченного набора токенов – словоупотреблений. Каждому словоупотреблению соответствует определенное слово, заданное в конкретном месте текста в определенной морфологической форме – словоформе. Каждому слову в языке соответствует набор словоформ, которые образуют морфологическую парадигму слова [1. С. 113]. Одна из словоформ парадигмы – лемма – задает слово в канонической форме (например, для сущ. – им. п., ед. ч; для прил. – им. п, ед. ч, муж. род; для глаголов – форма инфинитива и т.д.). Каждой словоформе соответствует набор морфологических признаков, состав и возможные значения которых зависят от части речи лексемы. Примеры словоформ: подводишь (Г ед, нст, 2л, дст), пол'янам (С мн, дт, жр), помен'явшем (ПРИЧАСТИЕ ед, пр, мр, прш, од, но, дст).

Целью морфологического анализа является сопоставление каждого словоупотребления в тексте с его словоформой, т.е. определение парадигмы слова, леммы слова, конкретной словоформы, а также набора признаков, которыми характеризуется данная словоформа (таких, например, как падеж, род и число существительного, время глагола и пр.). Например, словоформе **зодчеством** соответствует парадигма: з'одчество (С ед, им, ср)

[з'одчество (С ед, им, ср), з'одчества (С ед, рд, ср), з'одчеству (С ед, дт, ср), з'одчество (С ед, вн, ср), з'одчеством (С ед, тв, ср), з'одчестве (С ед, пр, ср), з'одчества (С мн, им, ср), з'одчеств (С мн, рд, ср), з'одчествам (С мн, дт, ср), з'одчества (С мн, вн, ср), з'одчествами (С мн, тв, ср), з'одчествах (С мн, пр, ср)].

Исторически морфологический анализ для языков с развитой морфологией, таких как русский, выполнялся двумя разными способами. Первый способ использует набор правил, в соответствии с которыми определяется словоформа. В простейших реализациях это таблица окончаний, по которой ищется парадигма. Одним из преимуществ такого подхода является то, что словоформу можно определить для любого произвольного слова языка. Существенный недостаток данного подхода в том, что язык содержит слишком много исключений, чтобы его можно было описать регулярными правилами. Поэтому в современных морфологических анализаторах данный подход используется только как вспомогательный для предсказания морфологической парадигмы неизвестных слов, которых нет в словаре анализатора.

Второй способ реализуется с использованием словаря, в котором описаны словоформы и соответствующие им морфологические признаки. Этот способ наиболее точный и быстрый по сравнению с предыдущим, особенно для такого морфологически развитого языка, как русский. Чем больше покрытие словарем слов языка, тем точнее анализ. В современных морфологических анализаторах русского языка этот способ является основным [2. С. 118]. В качестве словаря в подавляющем большинстве случаев используется словарь А.А. Зализняка [3] или его расширения [4]. Словарь представлен в виде набора морфологических парадигм, каждая из которых соответствует определенному слову – лексеме русского языка. Парадигма определяет список словоформ и лемму слова. Каждой словоформе соответствует список морфологических признаков. Словарь проиндексирован по словоформам и при нахождении соответствия статьи словаря определенной словоформе можно сразу получить всю парадигму данного слова, лемму, а также морфологические признаки искомой словоформы. Для упрощения реализации словаря каждая парадигма содержит псевдооснову и ссылку на словоизменительный класс. Словоизменительный класс является шаблоном, задающим набор псевдофлексий и связанных с ними словоизменительных признаков. По псевдооснове и словоизменительному классу строятся конкретные словоформы каждой парадигмы. В русском языке существует более полутора тысяч словоизменительных классов, которые описывают словоизменение любого слова русского языка. Пример словоизменительного класса для глагола **помен'ять** (выдержка): [*ть[ИНФИНИТИВ дст], *л[Г ед, мр, прш, дст], *ла[Г ед, жр, прш, дст], *ло[Г ед, ср, прш, дст], *ли[Г мн, прш, дст], *ю[Г ед, буд, 1л, дст], *ем[Г мн, буд, 1л, дст], *ешь[Г ед, буд, 2л, дст], *ете[Г мн, буд, 2л, дст], *ет[Г ед, буд, 3л, дст], *ют[Г мн, буд, 3л, дст], *в[ДЕЕПРИЧАСТИЕ прш, дст], *вши[ДЕЕПРИЧАСТИЕ прш, дст]...

Одним из недостатков использования словаря при определении морфологической парадигмы является то, что словарь содержит не все слова русского языка. Поэтому таким образом нельзя выполнить морфологический анализ тех слов, которые отсутствуют в словаре. Для подобных слов необходим тот или иной механизм предсказания их морфологической парадигмы.

В статье мы опишем один из методов такого предсказания – метод инкрементального предсказания парадигмы неизвестных слов, который позволяет с большой скоростью и высоким качеством выполнять предсказание. Одной из особенностей метода является то, что он работает в режиме самообучения полностью автоматически (unsupervised), постоянно наращивая качество предсказания по мере поступления новых слов (например, являясь одним из компонентов системы классификации документов или системы полнотекстового поиска).

Реализация

Инкрементальное предсказание морфологии неизвестных слов реализовано на базе следующих компонент – стратегий предсказания морфологии по единичной словоформе, лемматайзера и стратегии инкрементального предсказания морфологии по корпусу слов. Рассмотрим подробнее каждый из них.

Все стратегии предсказания основаны на морфологическом словаре А.А. Зализняка. Данный морфологический словарь представляет собой набор морфологических парадигм. Каждая морфологическая парадигма задает лемму и набор словоформ определенного слова, а также набор морфологических признаков каждой словоформы (род, число и пр.) и всей парадигмы в целом (часть речи и др.). Каждая парадигма слова связана с определенной флексивной моделью, определяющей словоизменительный класс парадигмы и задающей правила формирования словоформ парадигмы. Флексивная модель состоит из набора флексивных форм, каждая из которых задает псевдофлексию и морфологические признаки словоформ. Морфологический словарь содержит больше тысячи флексивных моделей, которыми исчерпывается все словоизменение современного русского языка. Морфологический словарь используется как для определения морфологической парадигмы уже имеющейся в нем словоформы (поиск на точное совпадение словоформы), так и для предсказания тех словоформ, которые явно не содержатся в словаре. Пример статьи морфологического словаря: фрезер'овка(С ед, им, жр) [фрезер'овка(С ед, им, жр), фрезер'овки(С ед, рд, жр), фрезер'овке(С ед, дт, жр), фрезер'овку(С ед, вн, жр), фрезер'овкой(С ед, тв, жр), фрезер'овкою(С ед, тв, жр), фрезер'овке(С ед, пр, жр), фрезер'овки(С мн, им, жр), фрезер'овок(С мн, рд, жр), фрезер'овкам(С мн, дт, жр), фрезер'овки(С мн, вн, жр), фрезер'овками(С мн, тв, жр), фрезер'овках(С мн, пр, жр)].

Для упрощения понимания алгоритмической реализации введем термины псевдопрефикс, псевдооснова, псевдосуффикс и псевдофлексия. Под

псевдопрефиксом слова будем понимать последовательность букв слова, которая начинается в начале слова и имеет заданный размер. Под псевдофлексией слова будем понимать последовательность букв слова, которая заканчивается в конце слова и совпадает с псевдофлексией одной из морфологических парадигм морфологического словаря. Под псевдосуффиксом слова будем понимать последовательность букв слова заданного размера, которая располагается непосредственно перед псевдофлексией. Под псевдоосновой слова будем понимать последовательность букв слова, которая остается, если убрать псевдопрефикс, псевдосуффикс и псевдофлексию. Более формально слово имеет следующий состав:

[<псевдопрефикс>] <псевдооснова> [<псевдосуффикс>]
[<псевдофлексия>],

Здесь знаками [...] показаны необязательные части слова, которые могут отсутствовать в конкретных словах. Данные термины введены только для понимания конкретных алгоритмов работы предсказания морфологической парадигмы и могут не совпадать с реальными префиксом, основой, суффиксом и флексией слова.

Стратегии предсказания по единичной словоформе пытаются предсказать морфологическую парадигму по единичной словоформе, без учета статистики, полученной по уже обработанным словоформам. Нами реализованы две такие стратегии – стратегия предсказания по псевдопрефиксу словоформы и стратегия предсказания по псевдосуффиксу словоформы. Стратегии базируются на алгоритмах, реализованных в проекте АОТ [5. С. 559].

Стратегия предсказания по псевдопрефиксу словоформы работает по следующему алгоритму (Алгоритм 1):

1. На основе словаря псевдопрефиксов пытаемся выделить из словоформы псевдопрефикс. Размер псевдопрефикса должен лежать в диапазоне [1, maxPrefixLength], при этом размер оставшейся части должен быть не меньше minSuffixLength.

2. Находим часть словоформы без псевдопрефикса и пытаемся предсказать ее парадигму на основе словаря морфологии (точное предсказание морфологии).

3. Если парадигма слова без псевдопрефикса найдена, формируем на ее основе новую парадигму добавлением ко всем словоформам исходной парадигмы заданного псевдопрефикса.

4. Рассматриваются только продуктивные части речи (существительные, прилагательные, глаголы и наречия).

Данная стратегия позволяет предсказывать парадигмы слов, образованных добавлением псевдопрефикса к уже существующему слову. Например, суперчеловек (супер + человек), псевдонаучный (псевдо + научный), антиправительственный (анти + правительственный), двуединный (дву + единый), антибарионы (анти + барион), вицеадмиральство (вице + адмиральство) и т.д.

Стратегия предсказания по псевдосуффиксу словоформы работает по следующему алгоритму (Алгоритм 2):

1. По всем словоформам морфологического словаря строится набор псевдосуффиксов, по которым будет выполняться предсказание. Псевдосуффикс включается в набор только от продуктивных частей речи и если размер словоформы не менее заданного `minBuildSuffixLength`. Кроме того, частотность флективной модели словоформы, по которой строится индекс, не должна быть менее заданной `minFlexiaModelFrequency`. Тем самым мы исключаем рассмотрение редких псевдосуффиксов

2. По исходной словоформе, парадигму которой требуется предсказать, пытаемся найти псевдосуффикс, совпадающий с псевдосуффиксом из набора, построенного выше. Псевдосуффикс должен быть в диапазоне [`minMatchedSuffixLength`, `minBuildSuffixLength`]. Поиск идет от наибольшего к наименьшему псевдосуффиксу.

3. Среди найденных словоформ по каждой части речи (если их более одной) находим ту парадигму, флективная модель которой является наиболее частотной. Остальные игнорируем. Это позволяет сильно сократить общее количество предсказанных словоформ.

Данная стратегия позволяет предсказывать парадигмы слов, имеющих неизвестную псевдооснову, но чье словоизменение подчиняется определенной флективной модели и чей псевдосуффикс известен морфологическому словарю. Например, дельта-антиген (-ген), диагностико-аналитический (-литический), думец-прогрессист (-прогрессист), варкалось (-алось), хликий (-ивкий), шорьки (-рьки), гуглить (-глить) и т.д.

Лемматайзер реализует составную стратегию предсказания, на вход которой приходит заданная словоформа, а на выходе, используя поиск на точное соответствие в морфологическом словаре и описанные выше стратегии предсказания, выдается набор парадигм словоформы, а также указание на конкретную словоформу, соответствующую исходной, в каждой парадигме. Лемматайзер работает по следующему алгоритму (Алгоритм 3):

1. Исходная словоформа ищется в морфологическом словаре на точное соответствие. Для поиска используется структура словоформ, построенная на базе DAWG (Directed Acyclic Word Graph) [6. Vol. 26, № 1. P. 3–16] – в нашей терминологии WordAutomat. WordAutomat позволяет компактно хранить все известные словоформы русского языка (порядка 5,5 млн) и эффективно осуществлять поиск по ним.

2. Если словоформа не найдена в морфологическом словаре, по каждой стратегии выполняется предсказание словоформ, после чего результаты предсказания объединяются.

Описанные выше компоненты выполняют предсказания по единичной словоформе и не учитывают результаты предсказания тех словоформ, которые уже были обработаны ранее. Поскольку все варианты словоизменения русского языка ограничены описанными выше флективными моделями, которые известны и уже имеются в морфологическом словаре, логично было бы предположить, что предсказание по нескольким словоформам может резко повысить качество предсказания. Для этого используется стратегия предсказания по корпусу слов. Стратегия работает в инкрементальном режиме.

тальном режиме (по мере поступления неизвестных слов) по следующему алгоритму (Алгоритм 4):

1. На вход стратегии предсказания по корпусу слов поступает очередная словоформа, морфологическую парадигму которой требуется предсказать.

2. Стратегия выполняет поиск парадигмы в морфологическом словаре на точное соответствие. Если парадигма (одна или несколько) найдена, то она используется в качестве результата.

3. Иначе стратегия предсказания проверяет, есть ли данная словоформа среди уже предсказанных. Если для словоформы уже было выполнено предсказание морфологической парадигмы ранее, используется ранее предсказанный вариант.

4. В противном случае стратегия предсказания делегирует работу по нахождению морфологической парадигмы стратегиям предсказания по единичной словоформе лемматайзера. Если лемматайзер вернул одну или несколько предсказанных морфологических парадигм, стратегия по найденным парадигмам пытается выполнить предсказание (см. далее процесс предсказания) и возвращается к п. 3 (п. 4 в этом случае не выполняется для предотвращения заикливания).

5. Если парадигмы для исходной словоформы не найдены, используется стратегия коррекции ошибок, которая ищет в морфологическом словаре все словоформы, расстояние Левенштейна [7. Т. 163; 4. С. 845–848] которых от заданной словоформы менее заданного значения (как правило, 1). Это позволяет исправить имеющиеся ошибки / опечатки в словах, корректируя заданное количество букв (одну или несколько).

Сам процесс предсказания инкрементальной стратегии предсказания по корпусу слов заключается в следующем (Алгоритм 5):

1. Стратегия ведет список частично предсказанных парадигм. Для каждой частично предсказанной парадигмы указываются псевдооснова (общая для всех словоформ часть), номер флективной модели из морфологического словаря и набор номеров флективных форм этой флективной модели, соответствующих обработанным словоформам.

2. По каждой предсказанной лемматайзером (его стратегиями предсказания по единичной словоформе) парадигме получают псевдооснова и флективная модель и индекс флективной формы в этой модели, которые сохраняются в частично предсказанных парадигмах.

3. Если количество различных индексов флективных форм определенной частично предсказанной парадигмы превысит заданный порог (по умолчанию 4), считаем, что парадигма корректно предсказана, удаляем ее из списка частично предсказанных и добавляем в WordAutomat полностью предсказанных парадигм. В дальнейшем она будет использована для предсказания соответствующих ей словоформ по алгоритму, описанному выше (Алгоритм 4, п. 3).

4. Большие корпуса текстов по закону Зипфа [8. Р. 484–490, 573] могут содержать большое количество редко используемых словоформ. Поэтому,

чтобы количество частично предсказанных парадигм не увеличивалось неограниченно, стратегия имеет LRU (Least Recently Used) очередь, в которой все частично построенные парадигмы выстроены в порядке их использования, с наиболее часто используемыми впереди очереди. При превышении размера очереди (по умолчанию, 10000) наиболее старые по времени использования парадигмы удаляются.

Примеры работы стратегии:

Базовые словоформы: дуршлак (С ед, им, мр), дуршлака (С ед, рд, мр), дуршлаке (С ед, пр, мр), дуршлагами (С мн, тв, мр).

Предсказанная парадигма: дуршлак (С ед, им, мр) [дуршлак (С ед, им, мр), дуршлака (С ед, рд, мр), дуршлаку (С ед, дт, мр), дуршлак (С ед, вн, мр), дуршлаком (С ед, тв, мр), дуршлаке (С ед, пр, мр), дуршлаки (С мн, им, мр), дуршлаков (С мн, рд, мр), дуршлакам (С мн, дт, мр), дуршлаки (С мн, вн, мр), дуршлагами (С мн, тв, мр), дуршлагах (С мн, пр, мр)].

Стратегия предсказания по корпусу текстов работает в полностью инкрементальном режиме автоматически, без участия человека (unsupervised). Это позволяет использовать ее в различных потоковых системах обработки, например в системах классификации документов, поисковых системах, диалоговых системах, где морфологический компонент системы часто может испытывать трудности при определении морфологии неизвестных слов. Подключение данной стратегии позволяет полностью прозрачно, т.е. без изменения исходной системы, наращивать качество предсказания морфологических парадигм за счет простой обработки все новых и новых словоформ. Причем эффективность используемых алгоритмов предсказания практически не снижает производительность всей системы в целом – не требуются громоздкие процедуры предобработки и предпостроения морфологических словарей по новым словоформам.

Исследование и результаты

В исследовании был использован корпус текстов, представляющий собой обезличенную коллекцию книг различной тематики на русском языке и содержащий более 10 млрд словоупотреблений¹.

В исследовании использовался морфологический словарь, построенный на базе морфологического словаря А.А. Зализняка и имеющий 5,5 млн словоформ и 174 тыс. парадигм².

Исследование проводилось в два этапа. На первом этапе анализировался корпус русского языка с учетом частотности распределения полученных словоформ. Поскольку учет частотности довольно затратен как по вычислительным ресурсам, так и по памяти, была проанализирована только часть полного корпуса размером 1 млрд словоупотреблений. Анализировались только знаменательные части речи. Для учета только нарицательных

¹ http://panchenko.me/data/russe/librusec_fb2.plain.gz

² <http://www.aot.ru/docs/rusmorph.html>

слов русского языка были отфильтрованы все словоупотребления, содержащие символы, отличные от букв алфавита русского языка (кроме дефиса), и все словоупотребления, начинающиеся на заглавную букву. Остальные словоупотребления были обработаны на основе описанной ранее стратегии инкрементального предсказания морфологии.

Методика расчетов параметров исследования выполнялась следующим образом:

– **Общее количество словоупотреблений.** Подсчитывалось каждое встреченное в тексте словоупотребление, участвующее в анализе.

– **Количество словоформ текста.** Формировался набор словоформ текста (Set). Если очередное анализируемое слово не содержалось в этом наборе, слово добавлялось в набор. В противном случае слово игнорировалось. По завершении анализа набор содержал список всех найденных словоформ текста. Размер данного списка соответствовал количеству словоформ текста.

– **Количество словоупотреблений, описываемых морфологическим словарем.** Для очередного анализируемого слова выполнялся поиск на точное соответствие в морфологическом словаре. Если словоформа, соответствующая данному слову, была найдена в словаре, слово подсчитывалось, в противном случае – нет. Для анализа распределения словоупотреблений по частям речи по каждой анализируемой части речи были сформированы свои счетчики, которые подсчитывали слова с парадигмой, соответствующей этой части речи.

– **Количество словоформ, описываемых морфологическим словарем.** Формировался набор словоформ, описываемых морфологическим словарем (Set). Для очередного анализируемого слова выполнялся поиск на точное соответствие в морфологическом словаре. Если словоформа, соответствующая данному слову, была найдена в словаре и слово в наборе отсутствовало, слово добавлялось в набор, в противном случае – нет. По завершении анализа набор содержал список всех найденных словоформ, для которых есть описание в морфологическом словаре. Размер данного списка соответствовал количеству словоформ, описываемых морфологическим словарем. Для анализа распределения словоформ по частям речи по каждой анализируемой части речи были сформированы свои наборы, куда попадали слова с парадигмой, соответствующей этой части речи.

– **Количество словоупотреблений с предсказанной парадигмой.** Для очередного анализируемого слова, которое не найдено в морфологическом словаре, выполнялось предсказание морфологической парадигмы. Если предсказание было выполнено успешно, слово подсчитывалось, в противном случае – нет. Для анализа распределения словоупотреблений по частям речи по каждой анализируемой части речи были сформированы свои счетчики, которые подсчитывали слова с парадигмой, соответствующей этой части речи.

– **Количество предсказанных парадигм.** Формировался набор предсказанных парадигм (Set). Для очередного анализируемого слова, которое

не найдено в морфологическом словаре, выполнялось предсказание морфологической парадигмы. Если предсказание было выполнено успешно и парадигма в наборе отсутствовала, парадигма добавлялась в набор, в противном случае – нет. По завершении анализа набор содержал список всех предсказанных парадигм. Размер данного списка соответствовал количеству предсказанных парадигм. Для анализа распределения парадигм по частям речи по каждой анализируемой части речи были сформированы свои наборы, куда попадали слова с парадигмой, соответствующей этой части речи.

– **Количество неизвестных словоупотреблений.** Если очередное анализируемое слово не было найдено в морфологическом словаре и предсказание морфологической парадигмы для него не было выполнено, слово подсчитывалось.

– **Количество неизвестных словоформ.** Формировался набор неизвестных словоформ (Set). Если очередное анализируемое слово не было найдено в морфологическом словаре, предсказание морфологической парадигмы для него не было выполнено и слово в наборе отсутствовало, слово добавлялось в набор. В противном случае слово игнорировалось. По завершении анализа набор содержал список всех неизвестных словоформ текста. Размер данного списка соответствовал количеству неизвестных словоформ.

– **Количество слов с опечатками (1 опечатка) из числа неизвестных словоформ.** Формировался набор словоформ с опечатками (Set). Если очередное анализируемое слово попало в категорию неизвестных, а стратегия коррекции ошибок нашла в морфологическом словаре словоформу, расстояние Левенштейна от которой до текущей парадигмы равно 1 и словоформа в наборе отсутствовала, словоформа добавлялась в набор. В противном случае словоформа игнорировалась. По завершении анализа набор содержал список всех словоформ с одной опечаткой. Размер данного списка соответствовал количеству словоформ с одной опечаткой.

– **Правильность предсказания** (% от общего количества предсказанных парадигм). Для оценки качества предсказания случайным образом была выбрана каждая 1000-ая парадигма, правильность предсказания которой проверялась вручную. Далее был посчитан % правильности предсказания по формуле: $(\text{кол-во правильно предсказанных парадигм}) / (\text{общее кол-во проверенных вручную парадигм}) * 100$.

– **Количество правильно предсказанных парадигм** было рассчитано методом аппроксимации по формуле: $(\text{общее кол-во предсказанных парадигм}) * (\text{правильность предсказания в \%}) / 100$.

– **Количество неправильно предсказанных парадигм** было рассчитано по формуле: $(\text{общее кол-во предсказанных парадигм}) - (\text{кол-во правильно предсказанных парадигм})$.

– **Количество словоформ с предсказанной парадигмой.** Формировался набор словоформ с предсказанной парадигмой (Set). Если для очередной анализируемой словоформы было успешно выполнено предсказание мор-

фологической парадигмы и словоформа в наборе отсутствовала, словоформа добавлялась в набор. В противном случае словоформа игнорировалась. По завершении анализа набор содержал список всех словоформ с предсказанной парадигмой. Размер данного списка соответствовал количеству словоформ с предсказанной парадигмой.

Общие параметры проанализированной части корпуса и результаты предсказания представлены в табл. 1.

Т а б л и ц а 1

Общие параметры частичного корпуса и результаты предсказания

Общий размер частичного корпуса (словоупотреблений)	1 082 797 004
Количество словоформ	3 525 863
Количество словоупотреблений, описываемых морфологическим словарем	924 764 959 (85,405%)
Количество словоформ, описываемых морфологическим словарем	124 876 (3,542%)
Количество словоупотреблений с предсказанной парадигмой	7 973 441 (0,736%)
Количество предсказанных парадигм	97 161 (2,756%)
Количество правильно предсказанных парадигм	93177
Количество неправильно предсказанных парадигм	3984
Правильность предсказания (% от общего количества предсказанных парадигм)	95.9%
Количество неизвестных словоупотреблений	150 058 604 (13,858%)
Количество неизвестных словоформ	3 303 826 (93,703%)

Распределение словоупотреблений по частям речи представлено в табл. 2–4.

Т а б л и ц а 2

Распределение словоупотреблений, описываемых морфологическим словарем, по частям речи

Сущ.	Прил.	Глаг.	Нареч.	Мест.-сущ.	Мест.-прил.
277 456 627 (25,624%)	85 764 441 (7,921%)	200 181 054 (18,487%)	24 744 417 (2,285%)	42 651 116 (3,939%)	38 474 087 (3,553%)

Т а б л и ц а 3

Распределение словоупотреблений с предсказанной парадигмой по частям речи

Сущ.	Прил.	Глаг.
4 689 994 (0,433%)	2 347 341 (0,217%)	917 681 (0,085%)

Т а б л и ц а 4

Распределение словоформ с предсказанной парадигмой по частям речи (в кол-ве парадигм)

Сущ.	Прил.	Глаг.
40 227 (1,141%)	43 765 (1,241%)	12 898 (0,366%)

Как видно из результатов, в словообразовании участвуют существительные, прилагательные и глаголы. Другие части речи, в частности наре-

чия, практически не представлены в данном корпусе в качестве новых слов. Скорее всего, это связано с тем, что наречия являются неизменяемыми, из-за чего предсказание их морфологической парадигмы описанным выше алгоритмом по нескольким словоформам работает плохо, так как не накапливается достаточного количества словоформ для формирования полной морфологической парадигмы.

На втором этапе был проанализирован полный корпус текстов без учета частотности. В процессе предсказания были также проанализированы слова, классифицированные как неизвестные, т.е. не описанные словарем анализатора и не попавшие в те, чья морфологическая парадигма была предсказана анализатором. Среди них как минимум четверть слов – это слова с одной опечаткой (по словарю анализатора). Можно предположить, что большая часть оставшихся слов – это слова с опечатками из числа тех, чья парадигма была предсказана, а также слова из морфологического словаря с большим (чем 1) количеством опечаток. Это предположение будет проверено в последующих исследованиях.

Были проанализированы словоформы при двух значениях параметра `minFlexiaFormsPerParadigm` – 4 и 7. Этот параметр определяет, при скольких заполненных словоформах частично построенная морфологическая парадигма считается предсказанной. Чем выше данный параметр, тем меньше будет количество ошибок предсказаний, но и меньшее количество словоформ будет предсказано. Как кажется, значение параметра 4 подходит для большинства практических случаев.

Результаты второго этапа исследования представлены в табл. 5–7.

Таблица 5

Параметры корпуса текстов

Размер корпуса (байт)	148Gb
Время анализа	3,5 ч
Скорость анализа	~1 млн словоупотреблений/с
Количество словоупотреблений	10 млрд
Количество словоформ	10 218 872
Количество словоформ, описываемых морфологическим словарем	2 038 283 (19,94%)
Количество неизвестных словоформ	7 610 325
Из них количество слов с опечатками (1 опечатка)	1 970 553 (19,27%)

Таблица 6

Результаты предсказания на корпусе текстов (`minFlexiaFormsPerParadigm` = 7)

Количество предсказанных парадигм	193282
Количество словоформ с предсказанной парадигмой	1 309 510 (12,81%)
Количество правильно предсказанных парадигм	189996
Количество неправильно предсказанных парадигм	3286
Правильность предсказания (% от общего количества предсказанных парадигм)	98,3%

Т а б л и ц а 7

Результаты предсказания на корпусе текстов (minFlexiaFormsPerParadigm = 4)

Количество предсказанных парадигм	570 264
Количество словоформ с предсказанной парадигмой	2 608 547 (25,52%)
Количество правильно предсказанных парадигм	539469
Количество неправильно предсказанных парадигм	30795
Правильность предсказания (% от общего кол-ва предсказанных парадигм)	94,6%

Для оценки качества предсказания на каждом этапе случайным образом была выбрана каждая 1000-ая предсказанная парадигма, правильность предсказания которой проверялась вручную.

Примеры правильно предсказанных слов:

– внесимволический, внесимволического, внесимволическому, внесимволического...

– двухметроворостый, двухметроворостого, двухметроворостому, двухметроворостого...

– изоморфический, изоморфического, изоморфическому, изоморфического...

– легаровский, легаровского, легаровскому, легаровского...

– микропарсек, микропарсека, микропарсеку, микропарсек, микропарсеком...

– сладкозвучнейший, сладкозвучнейшего, сладкозвучнейшему, сладкозвучнейшего...

– регулировать, регулирую, регулируем, регулируешь, регулируете...

Примеры неправильно предсказанных слов:

– Базовые словоформы: бен-барзилаю, бен-барзилаем, бен-барзилай, бен-барзилая.

Предсказанная парадигма (выдержка): бен-барзиль, бен-барзилаю, бен-барзилаем, бен-барзилаешь, бен-барзилаете...

– канцлеров, канцлерова, канцлерову, канцлеров, канцлеровом, канцлерове...

– юшкин, юшкина, юшкиного, юшкиному, юшкину, юшкина, юшкиного...

Примеры слов с опечатками: биллутень, оять, пезависимый, петебурге, хирур.

Примеры слов, для которых предсказание не было выполнено: хловиса, хлоп-хлоп, хмм, несусветские, завитневича, гс, шелдон, хеллоу.

Основное количество ошибок при предсказании составляют неверные предсказания заимствованных слов и имен собственных (которые попали в список анализируемых за счет неверного написания со срочной буквы).

Как видно из приведенных выше результатов, анализатор позволяет с высокой скоростью в полностью автоматическом режиме предсказать морфологическую парадигму значительной доли неизвестных слов (до

четверти всех словоформ), а также найти опечатки в уже известных словах, что позволяет с высокой точностью определить правильную морфологическую парадигму у двух третей словоформ на корпусе текстов 10 млрд словоупотреблений. Вопрос оценки полного частотного покрытия указанных словоформ на таком корпусе текстов оставим для последующих исследований.

Заключение

В статье был описан новый метод предсказания морфологической парадигмы неизвестных слов русского языка. Метод позволяет в инкрементальном режиме, т.е. по мере поступления очередной словоформы, в полностью автоматическом режиме, без участия человека предсказать морфологическую парадигму слова. Метод основан на ансамблевом предсказании морфологической парадигмы по одной словоформе по различным признакам – точного соответствия словоформы словарю, псевдопрефиксу, псевдосuffixу слова и др. На основе первичной оценки строятся частичные парадигмы, из которых по мере заполнения формируется предсказанная парадигма неизвестного слова, которая, в свою очередь, будет использоваться для дальнейшего предсказания. Также при анализе выполняется дополнительная коррекция опечаток, что дает возможность еще больше снизить количество ошибок определения морфологической парадигмы.

Данный метод позволяет с высокой скоростью правильно определить морфологическую парадигму до двух третей всех словоформ на корпусе текстов 10 млрд словоупотреблений.

В дальнейших исследованиях мы попробуем ответить на вопрос, какое реальное частотное покрытие по всем словоупотреблениям корпуса текстов дает такое предсказание. Предварительные оценки показывают значение не менее 95% от всех словоупотреблений, участвующих в анализе.

Кроме того, требует исследования вопрос влияния введенной в Алгоритм 5 очереди, ограничивающей общее количество анализируемых парадигм заданным размером.

Также требует дополнительного исследования оставшаяся часть неизвестных слов: являются ли они словами, имеющими больше одной опечатки, либо это очень редко употребляемые слова, словоформ которых недостаточно, чтобы сформировать полноценную парадигму.

Литература

1. Плузьян В.А. Общая морфология: Введение в проблематику. М. : Едиториал. УРСС., 2003. С. 113.
2. Клышинский Э.С. Начальные этапы анализа текста // Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие. М., 2011. С. 118.
3. Зализняк А.А. Грамматический словарь русского языка. М. : Русский язык, 1980.
4. AOT. URL: <http://aot.ru/docs/rusmorph.html>
5. Сокирко А.В. Морфологические модули на сайте www.aot.ru // Труды международной конференции «Диалог-2004. Компьютерная лингвистика и интеллектуальные технологии». М., 2004. С. 559.

6. Daciuk J. et al. Incremental construction of minimal acyclic finite-state automata // Computational linguistics. 2000. Vol. 26, № 1. P. 3–16.

7. Левенштейн В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов // Доклады Академии наук СССР. 1965. Т. 163, № 4. С. 845–848.

8. Zipf G.K. Human Behavior and the Principle of Least Effort. Addison-Wesley Press, 1949. P. 484–490, 573.

The Incremental Prediction of the Morphological Paradigm of Unknown Words in the Russian Language

Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya – Tomsk State University Journal of Philology. 2020. 68. 27–41. DOI: 10.17223/19986645/68/2

Elena V. Lyukina, Maria A. Lytaeva, Higher School of Economics (Moscow, Russian Federation). E-mail: eliouki@mail.ru / lytaeva_ma@mail.ru

Keywords: morphology, morphological analysis, morphological paradigm, prediction of morphological paradigm of unknown words, computational linguistics, corpus research.

This article describes a new method of prediction of the morphological paradigm of unknown (which are not in a dictionary) words in the Russian language. Modern morphological analyzers detect the morphological paradigm of a word using a dictionary of word forms, in which each word form has corresponding morphological features. This method is the fastest and the most precise in comparison with others, but has one essential shortcoming – it is limited to the available dictionary and cannot detect a morphological paradigm of unknown words, i.e. those words not given in the dictionary. The method described in the article allows, in the incremental mode, predicting the morphological paradigm of a word. The method is based on the ensemble prediction of the morphological paradigm by a single word form and the consecutive formation of partial paradigms by several word forms. Partial paradigms then are used to compute final prediction. At the first step, ensemble prediction polls several various prediction strategies and forms an intermediate result of prediction. At the second step, the method correlates collected word forms and builds partial paradigms. The partial paradigms, which are filled by word forms to some threshold, then are used to form final result of prediction. At the third step, error correction and new prediction are performed for words whose morphological paradigm cannot be detected. An important advantage of the described method of prediction is that it works in the incremental unsupervised mode, without human intervention. The system is self-learning – the more word forms are met, the quicker and more precise the result of prediction is. Also, prediction algorithms practically do not affect the performance of the overall system. Bulky procedures of preprocessing and precreation of morphological dictionaries for new word forms are not required. To confirm the applicability of the new method, a research was conducted. The research used a text corpus of the Russian language of different genres (approx. 10 bn words). The research was done in two steps. At the first step, the analysis of a partial text corpus (about 1 bn words) was performed. The analysis took into account the frequency distribution of the predicted word forms. Only nouns, adjectives, verbs and adverbs were analyzed. At the second step, the full text corpus was analyzed without frequency distribution. The research confirmed the high precision and performance of the described prediction method. The method is able to predict morphological paradigms of two thirds of all word forms met in the text.

References

1. Plungyan, V.A. (2003) *Obshchaya morfologiya. Vvedenie v problematiku* [General morphology. Introduction to the problem field]. Moscow: Editorial. URSS.

2. Klyshinskiy, E.S. (2011) Nachal'nye etapy analiza teksta [Initial stages of text analysis]. In: Bol'shakova, E.I. et al. *Avtomaticheskaya obrabotka tekstov na estestvennom yazyke i komp'yuternaya lingvistika: ucheb. posobie* [Automatic processing of texts in natural language and computational linguistics: textbook]. Moscow: MIEM.

3. Zaliznyak, A.A. (1980) *Grammaticheskiy slovar' russkogo yazyka* [Grammar dictionary of the Russian language]. Moscow: Russkiy yazyk.
4. AOT. [Online] Available from: <http://aot.ru/docs/rusmorph.html>.
5. Sokirko, A.V. (2004) [Morphological modules on the site www.aot.ru]. *Dialog-2004. Komp'yuternaya lingvistika i intellektual'nye tekhnologii* [Dialogue-2004. Computational Linguistics and Intelligent Technologies]. Proceedings of the International Conference. Moscow: Nauka. (In Russian).
6. Daciuk, J. et al. (2000) Incremental construction of minimal acyclic finite-state automata. *Computational Linguistics*. 26 (1). pp. 3–16.
7. Levenshteyn, V.I. (1965) Dvoichnye kody s ispravleniem vypadeniy, vstavok i zameshcheniy simvolov [Binary Codes with Correction of Dropouts, Insertions and Substitutions of Symbols]. *Doklady Akademiy Nauk SSSR*. 163.4. pp. 845–848.
8. Zipf, G.K. (1949) *Human Behavior and the Principle of Least Effort*. Addison-Wesley Press. pp. 484–490.