УДК 81'322

АВТОМАТИЗАЦИЯ ГЕНЕРАЦИИ МОРФОЛОГИЧЕСКИХ ФОРМ РУССКИХ ПРЕДИКАТОВ

О.И. Бабина, П.Г. Осминин

Статья описывает модель морфологической генерации словоформ, сочетающую словесно-парадигматический и элементно-операционный подходы. Представлен обзор базы знаний и алгоритма автоматизации морфологической генерации предикатов. Даны сведения о случаях, не покрываемых моделью.

Ключевые слова: морфологическая генерация, предикат, русский язык, флексия, словарь.

1. Введение

Уровень морфологического анализа и синтеза в системах обработки естественного языка – минимальный уровень лингвистического анализа. Лексемы представлены в тексте в своих парадигматических формах, каждая из которых характеризуется набором морфологических признаков, характерных для части речи соответствующей лексемы. Определение морфологических характеристик слов в тексте реализуется в системах обработки естественного языка путем разметки на части речи и является основой для последующего лингвистического анализа.

Для представления морфологического уровня лексической базы знаний существуют следующие подходы:

1) в виде морфологических лексиконов, состоящих из минимально требуемых лексических единиц (слова, морфемы, квази-морфемы) и алгоритма, проводящего морфологическую генерацию или анализ словоформ, составленных из лексиконов. Выделяются две группы подходов к морфологической генерации: элементно-комбинаторный (Item-and-Arrangement) и элементно-операционный (Item-and-Process). В подходах первой группы слово рассматривается как линейная последовательность (квази)морфем, примеры работ в этом направлении фреймовый подход [1], аддитивная модель [2] и др. В подходах второй группы морфемный изменения обуславливаются определенными морфологическими процессами. Примеры работы в этом направлении генеративная модель [3] и др.;

2) в виде полного списка «готовых» словоформ, имеющих набор грамматических характеристик, представленных разметкой по частям речи [4]. Такой подход к организации лексикона следует старейшей модели грамматического описания – словесно-парадигматической (Word-and-Paradigm). В рамках такой модели лексема рассматривается как набор слово-

форм, имеющих общий корень, сгруппированных в общую парадигму. Главной единицей становится словоформа, а не морфема.

У каждого подхода есть свои преимущества и недостатки. Автоматическая генерация морфологических парадигматических форм универсальна и способна покрыть морфологию словоформ, не представленных эксплицитно в лексиконе, но в языках флективного типа довольно сложно формализовать морфологическую генерацию. Ограничение предметной области текстов может значительно сократить количество морфологических форм, необходимых системе обработки естественного языка, что дает возможность создать лексическую базу знаний с высоким процентом покрываемости, достаточной для надежной работы системы обработки языка.

Учитывая вышесказанное, мы используем преимущества обоих подходов при построении лексикона. В нашей модели лексическая база знаний должна иметь морфологическую зону для каждого вхождения. Морфологическая зона должна заполняться текстовыми словоформами, отражающими морфологические варианты соответствующей лексемы; варианты содержатся в лексиконе в виде готовых словоформ [5, 6, 7]. Таким образом, лексема представлена морфологическими парадигматическими формами.

Недостатком такого подхода является трудоемкость ручного заполнения парадигматических полей в морфологической зоне словаря. Мы предлагаем автоматизировать морфологическую генерацию парадигматических форм на стадии составления лексикона с обязательной проверкой сгенерированной парадигмы человеком. Преимущества такого подхода очевидны – экономия времени на стадии заполнения базы знаний и отсутствие ошибок в базе знаний, и следовательно, при работе системы автоматической обработки текста на этапе парадигматической идентификации словоформ.

2. Автоматическая морфологическая генерация русских предикатов

2.1. Парадигма русских предикатов

Предикаты играют важную роль в формировании предложений. Независимо от своей принадлежности к части речи, предикаты характеризуются семантическими, функциональными и морфологическими свойствами. Предикаты – особый лексико-грамматический класс. Наше внимание концентрируется на формальном аспекте предикатов – морфологической парадигме, которая должна быть отражена в морфологической зоне лексикона.

После анализа корпуса русскоязычных научных текстов был выявлен список предикатов. Анализ показал, что большинство предикатов являются глаголами. Кроме глаголов предикаты могут выражаться полными и краткими прилагательными и краткими причастиями.

На основе формального, функционального и семантического критерия предикаты разделены на семантические классы активных и пассивных предикатов. Активный предикат это глагол, выполняющий функцию пре-

диката в предложении, а его формальное подлежащее (или управляющий член предложения) играет роль агента или темы (такое значение присуще активного залогу, что отражено в названии класса). У пассивного предиката формальное подлежащее в предложении играет роль пациенса. Анализ корпуса показал, что активные предикаты представлены как возвратными, так и невозвратными глаголами. В класс пассивных предикатов входят возвратные глаголы, прилагательные и краткие формы причастий. Таким образом, классы предикатов, в основном, разграничены формально, за исключением того, что оба класса включают в себя возвратные глаголы; последние могут быть разведены семантически. Проанализировав состав морфологических форм, образуемых активными и пассивными предикатами, мы пришли к выводу о различиях в составе морфологических парадигм двух классов. Так, только активные предикаты имеют форму деепричастия и могут использоваться в повелительном наклонении.

Корпусные анализ научных текстов дает следующие 6 групп форм морфологической парадигмы для предикатов (группы 1–5 соответствуют изъявительному наклонению):

1) личные формы (*зажат, идентична, идет, накопили, поворачивает*): характеризуются грамматическими категориями времени, лица и числа. В проанализированном корпусе представлены все времена (прошедшее, настоящее и будущее) и числа (единственное и множественное), характерные для русского языка. Категория лица представлена в основном третьим лицом, встречаются также формы первого лица. Формы второго лица отсутствуют;

2) полные формы (*затрачиваемой, излучающие, изобретенных, разрезанный*) – выражены прилагательными или (для глагольных предикатов) причастиями. Причастия обладают свойствами глаголов и прилагательных. Глагольные свойства проявляются через функциональный аспект – причастия функционируют как предикаты в определительных придаточных предложениях, частотных в научных текстах. По формальному критерию причастия аналогичны прилагательным – парадигматическая форма причастий определяется категориями числа (единственное и множественное), рода (мужской, женский, средний) и падежа (именительный, родительный, дательный, винительный, творительный, предложный);

3) инфинитив – неличная форма глагола (*закрепить, затянуть, изготовляться, направить*). Для предикатных прилагательных и кратких причастий инфинитив формируется как составное сказуемое с глаголом-связкой «быть», который используется в форме инфинитива. Семантически значимая часть составного сказуемого, выраженная прилагательным и причастием, изменяется по числам и родам (для единственного числа). Поэтому предикаты, выраженные прилагательными и причастиями, могут иметь более одной формы в инфинитиве;

4) модальные формы (*может вступать, может быть определено, должен показывать, могло приводить*). Составные глагольные предикаты, выраженные сочетанием инфинитива с личными формами в изъявительном наклонении модальных глаголов «может» или «должен» достаточно частотны в корпусе и они включаются в парадигму предикатов. Глагол «может» изменяется по числам, а глагол «должен» может изменяться по числам и родам (для единственного числа). В русскоязычных научных текстах, часто используются безличные предложения с предикатами, выраженными инфинитивом с модальным модификатором. Поэтому такие формы учитываются в этой группе;

5) деепричастие (*исполняя, нажимая, обмениваясь, передавая*). В корпусе деепричастия использовались только в настоящем времени;

6) повелительные формы. Формы повелительного наклонения семантически подходят только для активных предикатов. В русскоязычных научных текстах такие формы строятся аналитически для несовершенного вида и в этом случае совпадают с личными формами первого лица множественного числа будущего времени (*будем называть, будем обозначать, будем округлять*); для совершенного вида формы повелительного наклонения образуются синтетически (*закрепим*, *исключим, рассмотрим*).

Из приведенного набора морфологических характеристик можно заметить, что некоторые предикаты (прилагательные и краткие причастия) отражают категорию рода, что характерно для именных частей речи. Перечисленные предикаты формируют группу пассивных предикатов, которую можно разделить на два подкласса – возвратные глаголы и прилагательные/краткие причастия. В зависимости от вида пассивного предиката некоторые формы могут не отражать категорию рода. Морфологические варианты предикатов в безличных предложениях (отражены в группе модальных форм) не характерны для возвратных глаголов, что обусловливает разницу в парадигме двух подклассов.

Анализ парадигм предикатов показывает, что формально предикаты могут быть классифицированы по спряжению. В традиционной грамматике выделяется только два класса спряжения, что недостаточно для формализации базы знаний для автоматической морфологической генерации. В следующем разделе мы покажем, что схемы спряжения применимы в качестве базы знаний для аддитивной модели автоматической генерации морфологической парадигмы русского предиката.

2.2. Классы спряжения русских предикатов

В качестве основной формы будем рассматривать форму причастия (для глагольных предикатов) или прилагательного в единственном числе мужском роде именительном падеже. В зависимости от схемы словообразования основные формы можно разделить на а) причастия несовершенного вида; б) причастия совершенного вида; в) прилагательное. Несовершенный вид причастий образуется с помощью суффиксов *-ущ/-ющ/-ящ*, на-

пример, «*влекущий*», «*комбинирующий*», «*водящий*» и т.д. Совершенный вид причастий формируется с помощью суффиксов *-вш/-ш*, например, «*посвятивший*», «*внесший*».

Парадигма активных предикатов представлена причастиями несовершенного вида, изменяющимися в соответствии с восемью схемами спряжения. Схемы формально отличаются набором окончаний, используемых в различных парадигматических формах. Для основной формы это следующие окончания: 1) *-ующий*; 2) *-аюший*; 3) *-ащий*; 4) *-ящий*; 5) *-ующийся*; 6) *-ающийся*; 7) *-ащийся*; 8) *-ящийся*.

Для пассивных предикатов мы выдели три группы со следующими основными формами: 1) причастие несовершенного вида, 2) причастие совершенного вида, 3) прилагательное. Перечисленные группы можно разделить на более дробные классы в соответствии с используемой схемой спряжения, что легко может быть формализовано в рамках элементно-комбинаторной модели. В группах выделяются следующие классы:

1. Причастия, несовершенный вид:

а) *-уемый*: «*наследуемый*», «*генерируемый*», «*интегрируемый*»;

б) *-емый*: «*предоставляемый*», «*предлагаемый*», «*повторяемый*»;

в) *-мый*: «*любимый*», «*зависимый*», «*проводимый*»;

г) *-ющийся*: «*пополняющийся*», «*анализирующийся*»;

д) *-ащийся*: «*содержащийся*»;

е) *-ящийся*: «*находящийся*», «*годящийся*», «*сводящийся*».

2. Причастия, совершенный вид:

а) *-нный*: «*пониженный*», «*вычисленный*», «*выявленный*»;

б) *-вшийся*: «*разрушившийся*», «*встретившийся*», «*выявившийся*».

3. Прилагательные:

а) *-ьный*: «*привлекательный*», «*максимальный*», «*оптимальный*»;

б) *-ный*: «*причастный*», «*ясный*», «*видный*»;

в) -[шипящий]*ий*: «*хороший*», «*свежий*»;

д) -[заднеязычный]*ий*: «*мягкий*», «*легкий*», «*близкий*»;

е) -[шипящий]*ой*: «большой»;

д) *-ой*: «*развитой*», «*простой*», «*таковой*».

Следует отметить, что формально классы возвратных глаголов идентичны семантическому классу активных предикатов, поэтому их схемы генерации одинаковы за исключением некоторых отсутствующих форм в пассивных предикатах (инфинитив, повелительное наклонение). Присутствие возвратных глаголов в активном и пассивном семантическом классе означает, что существуют случаи, когда установить границу между классами возможно только исходя из семантики глаголов, а не из формальных признаков. Таким образом, семантический класс (активный или пассивный) выбирается для таких предикатов вручную. Ранее выделенные классы прилагательных для именной генерации [8] были расширены для отражения личных форм, требуемых для генерации предикатов.

Таким образом, морфологическая генерация предикатных форм может быть легко формализована и автоматизирована. В следующем разделе мы представим структуру базы знаний и общий алгоритм морфологической генерации.

2.3. База знаний и алгоритм автоматической морфологической генерации

После анализа морфологических вариантов в выделенных схемах спряжений предикатов модель морфологической генерации в рамках элементно-комбинаторного подхода можно представить как: *Основа* + *Суффикс* + *Окончание* для синтетических форм и *Связка* + *Инфинитив* для аналитических форм (где *Связка* – фиксированная словоформа глагола-связки или модального глагола, *Инфинитив* генерируется синтетически). В нескольких схемах суффикс требуется для генерации определенных форм. В целях унификации модели мы считаем, что другие формы (генерируемые сочетанием корня и окончания) содержат нулевую квазиаффиксальную морфему. В такой схеме меняется только окончание.

Таким образом, база знаний, требуемая для алгоритма генерации, состоит из морфологических схем для каждого выделенного класса предикатов. Например, для класса 1 активных предикатов фрагмент схемы формообразования выглядит следующим образом:

Причастие, ед.ч., м.р., им.п.	Основа + *ующ* + *ий*
Личная форма, наст.вр., ед.ч.	Основа + *у* + *ет*
Личная форма, прош.вр., ед.ч., ж.р.	Основа + *ова* + *ла*
Инфинитив	Основа + *ова* + *ть*
Личная форма, буд.вр., ед.ч.	*будет* + Инфинитив

В базу знаний входит правило образования основы (путем опущения окончания) для каждого класса.

Алгоритм морфологической генерации предикатных форм состоит из следующих шагов (шаги, выполняемые вручную, оговорены):

1) вручную выбрать класс предиката (активный или пассивный);

2) определить морфологический класс предиката (сравнить со списком окончаний и выбрать класс с наибольшим по длине совпадающим окончанием);

3) определить основу (при помощи правила из базы знаний);

4) сгенерировать синтетические морфологические формы предиката в соответствии с морфологическими схемами из базы знаний;

5) вручную сгенерировать формы с нерегулярным формообразованием (если требуется);

6) сгенерировать аналитические морфологические формы предиката в соответствии с морфологическими схемами из базы знаний.

Таким образом, вручную выполняются только выбор семантического класса (шаг 1) и (для некоторых схем) генерация инфинитива (шаг 5).

3. Заключение

Экспериментальная оценка модели показала ее перспективность – большинство форм генерируется корректно. Однако из-за сложности естественного языка существуют случаи, которые не покрываются существующей моделью.

Во-первых, в модели не учитываются предикаты с чередующимися гласными в корне. Например, формы предиката «*кроющийся*» совпадают с классом 6 активных предикатов несовершенного вида. Личные формы в будущем и прошедшем времени, инфинитив содержат чередующую гласную в корне – «*будет крыться*» (буд.вр., ед.ч.), «*крылся*» (прош.вр., ед.ч., м.р.).

Во-вторых, в личных формах будущего времени и инфинитиве предикатов совершенного вида происходит чередование согласных. Например, полная форма предиката «*привлекший*» в личной форме множественного числа будущего времени имеет вид «*привлечем*» – происходит чередование букв *к – ч*.

В-третьих, модель не учитывает исключения из общих схем генерации морфологических форм. Например, предикат «*мажущий*» не имеет формы деепричастия. Такие случаи и их причины (формальные и семантические) требуют дальнейшего исследования.

Таким образом, предложенная модель морфологической генерации предикатов следует элементно-комбинаторному подходу и покрывает стандартные схемы спряжения глаголов и прилагательных, использующихся в функции предикатов. При этом хранение полной морфологической парадигмы лексем в лингвистической базе знаний соответствует словесно-парадигматическому подходу. Текущая модель не покрывает случаи неправильного синтеза форм с чередованием гласных или согласных, не имеющих полной парадигмы предикатов. Изучение перечисленных случаев является направлением дальнейшего исследования.

Библиографический список

1. Lukanin, A. Frame approach to Persian verb generation for educational purposes / A. Lukanin, C. Bobroff // Proceedings of the 2nd Workshop on Computational Approaches to Arabic Script-based Languages. – Stanford, California, USA: Linguistic Institute, 2007. Pp. 98–105.

2. Бабина, О.И. Нестрого аддитивный подход к автоматическому морфологическому анализу флективных языков / О.И. Бабина, Н.Ю. Дюмин // Материалы 5-й Междунар. науч.-практ. конф. «Наука и современность-2010». Секция «Филологические науки». – Новосибирск: Центр развития научного сотрудничества, 2010. – С. 12–17.

3. Siegel, D. Topics in English Morphology: Doctoral dissertation MIT / D. Siegel // Garland, New York, 1974.

4. Sheremetyeva, S. Natural Language Analysis of Patent Claims / S. Sheremetyeva // Proceedings of the ACL 2003 Workshop on Patent Corpus Processing. – Sapporo, Japan, 2003. Pp. 66–73.

5. Sheremetyeva, S. «Less, easier and quicker» in language acquisition for patent MT / S. Sheremetyeva // Proceedings of Workshop on Patent Translation, MT Summit X. – Phuket, Thailand, 2005. Pp. 35–42.

6. Бабина, О.И. Построение базы лингвистических знаний для многоязычных систем автоматической обработки текстов / О.И. Бабина // Лингвистика в контексте культуры: материалы V международной научно-практической конференции. – Челябинск: Изд. центр ЮУрГУ, 2012. – С. 19–23.

7. Шереметьева, С.О. К вопросу об электронных ресурсах профессиональной лексики / С.О. Шереметьева, П.Г. Осминин, Е.С. Щербаков // Вестник ЮУрГУ. Серия «Лингвистика». – 2014. – Т. 11. – № 1. – С. 57–63.

8. Babina, O. Modes of Automating Lexicon Compilation as a Component of Practical Studies for Linguists / O. Babina // General and Professional Education. – 2012. № 2. Pp. 3–12.