Математико-алгоритмическая формализация моделей морфологического анализа и синтеза словоформ естественных языков

А. В. Пруцков

Рязанский государственный радиотехнический университет
390005, Рязань, ул. Гагарина, 59/1

e-mail: mail@prutzkow.com

Аннотация. Модели морфологического анализа и синтеза словоформ разрабатывались на основе исследований морфологии естественных языков, то есть с использованием лингвистического подхода. Однако для использования этих моделей в информационных программных системах автоматической обработки текстов необходимо провести их анализ с математико-алгоритмической точки зрения. Целью статьи является формальное описание моделей морфологического анализа и синтеза словоформ и их анализ по различным аспектам. Аспектами являются способы алгоритмической реализации морфологического анализа и синтеза словоформ, используемые операции и их связи с объектами моделей, способы сокращения пространства поиска при морфологическом анализе. Анализируются элементно-комбинаторная, элементно-операционная и словесно-парадигматическая модели, а также предложенная универсальная модель формообразования. Математико-алгоритмическая формализация моделей выявила их схожесть. Сделан вывод, что во всех моделях, кроме элементно-комбинаторной, операции в явном виде не определены и чаще всего являются добавлением аффикса или замены части основы на другую. Рассмотренные модели уменьшают пространство поиска при морфологическом анализе.

Ключевые слова: автоматическая обработка текстов, морфологический анализ, морфологический синтез, элементно-комбинаторная модель, элементно-операционная модель, словесно-парадигматическая модель.

1. Лингвистический и математико-алгоритмический подходы к исследованию языковых процессов

«Из имеющихся в распоряжении человека интеллектуальных средств познания сложных явлений самым мощным является абстракция. Отправной момент абстракции заключается в фиксировании сходства некоторых объектов, ситуаций или явлений реального мира и решении выделить сходные свойства, временно оставляя в стороне различия» [1].

Любая предметная область представляет собой совокупность объектов и задач, которые в ней необходимо решать для достижения определенных целей. Решение задачи является процессом, который можно представить в виде алгоритма. Насто-

ящая статья посвящена процессам анализа и синтеза текста на естественных языках. Эти процессы можно рассматривать с различных точек зрения.

Изначально процессы в языке исследовались в лингвистике. Лингвистика создала фундаментальную научную базу описания таких процессов, включающую:

- модели процессов на основе выделенных свойств и особенностей;
- взаимосвязи процессов друг с другом;
- классификацию и группировку процессов на основе их особенностей в естественных языках одной или различных групп и семейств.

С увеличением аппаратной мощности компьютерной техники и разработкой информационных программных систем, использующих эти мощности, появилась возможность реализовывать модели процессов на компьютерах. Однако существующие лингвистический модели были нечетко формализованы и нуждались в адаптации или практически полной переработке. В результате для разработки моделей был использован математико-алгоритмический подход (например, формальные языки [2]). Модели, разработанные с помощью этого подхода, имеют высокую степень абстракции и могут быть реализованы в информационных программных системах.

Математико-алгоритмические модели имеют различные уровни формализации, однако их средний уровень выше, чем у лингвистических моделей. Достижение высокого уровня формализации затруднительно, так как естественный язык является неформализованным объектом (более подробно проблемы формализации естественного языка рассмотрены в работе [3]).

Далее в статье будут рассмотрены модели процесса морфологического анализа и синтеза словоформ, предложенные на основе обоих подходов.

2. Цель статьи

Целью статьи является математико-алгоритмическая формализация моделей морфологического анализа и синтеза словоформ на основе единого описания предметной области, анализ формализованных моделей и выявление их сходств и различий с математико-алгоритмической точки зрения.

Анализ моделей будет проводиться по следующим аспектам:

1. Способы реализации процессов морфологического анализа и синтеза словоформ с алгоритмической точки зрения.

2. Анализ используемых в моделях операций, их возможных видов и их связей с объектами, к которым они применяются.

3. Подходы к сокращению пространства поиска при морфологическом анализе. Задача морфологического анализа является переборной, поэтому для уменьшения временно́й сложности необходимо сократить пространство поиска.

Модели, которые будут рассмотрены в статье, могут быть применены в информационных программных системах автоматической обработки текстов. Поэтому их описание будет максимально возможно формальным и абстрактным.

3. Термины и обозначения

3.1. Лингвистические термины

Приведем определения лингвистических терминов, необходимых для дальнейшего изложения.

Лексема — «слово, рассматриваемое как единица словарного состава языка в совокупности его конкретных грамматических форм» [4].

Грамматическое значение — «значение, присущее ряду слов и находящее в языке свое регулярное (стандартное) выражение» [4].

Парадигма — система словоформ, соответствующих лексеме.

Морф — неделимая часть слова. Корневой морф будем называть основой. Аффиксальный морф будем называть аффиксом (в том числе и флексию).

Морфема — минимальная часть слова. Выражается одним или несколькими морфами.

Морфологический анализ словоформы (далее — морфологический анализ) — процесс определения грамматического значения словоформы и при необходимости выделения в ней основы, сопоставляемой с лексемой.

Морфологический синтез словоформы (далее — морфологический синтез) — процесс генерации словоформы с заданным грамматическим значением с использованием основы, сопоставленной с лексемой.

3.2. Обозначения

В описании моделей и задач морфологического анализа и синтеза будем использовать следующие обозначения.

$A, B, D, E, ..., Z$ — множество.

$AA, AB, AD, ..., ZZ, AAA, AAB, ..., ZZZ, ...$ — отношение множеств, обозначенных прописными латинскими буквами, или специально введенное отношение.

$a_i, b_i, ..., z_i$ — i-й элемент множества.

$aa_i, ab_i, ac_i, ..., zz_i, aaa_i, aab_i, ..., zzz_i, ...$ — i-й элемент отношения.

$a, b, ..., z, aa, ab, ac, ..., zz, aaa, aab, ..., zzz, ...$ — конкретный элемент множества.

$cA, cB, ..., cZ, cAA, cAB, ..., cZZ, cAAA, cAAB, cZZZ, ...$ — мощность множества, обозначенного прописными латинскими буквами.

caa, cab, ..., czz, caaa, caab, ..., czzz, ... — количество составляющих элемента отношения, обозначенного второй и последующими строчными латинскими буквами.

4. Морфологический анализ и синтез словоформ

4.1. Область применения морфологического анализа и синтеза словоформ

На морфологическом уровне автоматической обработки текстов выполняются морфологический анализ и синтез. Морфологический анализ и синтез словоформ используется при решении многих задач автоматической обработки текстов, среди которых стоит выделить следующие:

- анализ и оценка текстов коллективных договоров [5];
- интеллектуальный анализ текстов [6];
- поиск аналогов объектов по их текстовому описанию [7];
- анализ различных текстовых структур [8, 9];
- информационный поиск в коллекциях текстов [10];
- исследование других уровней автоматической обработки текстов [11];
- лингвистическое автоматизированное обучение [12].

4.2. Математическое описание предметной области

Предметная область включает следующие множества объектов.

$G = \{g_1, g_2, ..., g_{cG}\}$ — множество грамматических значений.

$V = \{v_1, v_2, ..., v_{cV}\}$ — множество граммем.

$L = \{l_1, l_2, ..., l_{cL}\}$ — множество лексем.

$W = \{w_1, w_2, ..., w_{cW}\}$ — множество словоформ.

Множества связаны отношением $cW \gg cL.$

Каждое грамматическое значение является множеством граммем

$$g_i \subset V, 1 \le i \le cG.$$

Множества связаны отношением $L \times G \times W.$

Для реального естественного языка $LGW \subset L \times G \times W.$

Введем требование, чтобы составляющая или сочетание составляющих элемента отношения не повторялись в других элементах, и будем обозначать неповторяющиеся составляющие отношения чертой сверху. Например, в элементах отно-

шения LGW не должно повторяться сочетание элементов l и g: $\overline{LGW}$, что означает:

$$\forall l\,\mathrm{g}\,w_i = (l_i, \mathrm{g}_i, w_i)\ \forall l\,\mathrm{g}\,w_j = (l_j, \mathrm{g}_j, w_j)\ ((i \neq j) \wedge (l_i \neq l_j) \wedge (g_i \neq g_j)),$$

$$1 \leq i \leq cLGW, 1 \leq j \leq cLGW,$$

где $\wedge$ — логическая операция конъюнкция.

Другими словами, составляющие l и g однозначно определяют элемент отношения $l\,\mathrm{g}\,w$.

4.3. Математическая постановка задачи морфологического анализа и синтеза

Введем операцию получения неизвестных составляющие элементов отношения по известным составляющим. Например, операция возвращает элементы u, y отношения $UXYZ$ по элементам x, z: $(u, y) \leftarrow UXYZ[x, z]$. Предполагается, что $U\overline{XY}\overline{Z}$.

Чтобы решить задачу морфологического синтеза, необходимо найти элемент $lgw = (l, g, w) \in LGW$, включающий заданные элементы l и g, и из него определить w (алгоритм 1).

Алгоритм 1. Обобщенный алгоритм морфологического синтеза

Вход: Лексема l, грамматическое значение g.

Выход: Словоформа w, соответствующая l и g.

1 $(w) \leftarrow LGW[l, g]$

2 **return** w

Пространство поиска при морфологическом синтезе невелико и не представляет проблемы при решении этой задачи.

Чтобы решить задачу морфологического анализа, необходимо найти элемент $lgw = (l, g, w) \in LGW$, включающий заданный элемент w, и из него определить l и g (алгоритм 2).

Алгоритм 2. Обобщенный алгоритм морфологического анализа

Вход: словоформа w.

Выход: лексема l, грамматическое значение g, соответствующие w.

1 **for** $(l_i, g_i, w_i) \in LGW, 1 \le i \le cLGW$ **do**
2 **if** $w == w_i$ **then**
3 $l = l_i$
4 $g = g_i$
5 **break**
6 **end if**
7 **end for**
8 **return** l, g

Алгоритм морфологического анализа связан с перебором всех словоформ лексем и сравнении их с исходной словоформой. Пространства поиска значительно, что влияет на временну́ю сложность алгоритма. Рассматриваемые далее модели реализуют различные подходы к сокращению пространства поиска.

Здесь и далее алгоритмы описываются с использованием множеств, отношений и операций с ними, введенными в статье.

4.4. Проблемы морфологического анализа и синтеза

Проблемы компьютерного морфологического анализа и синтеза можно разделить на алгоритмические и лингвистические.

Математическая постановка задачи предполагает хранение всех словоформ и использование их для морфологического анализа и синтеза. Преимуществом морфологического анализ и синтеза с использованием списка словоформ является простые модели этих процессов и алгоритмы (см. алгоритмы 1 и 2). Такой подход используется на практике [13]. Однако при большом числе словоформ морфологический анализ требует значительных временны́х затрат, так как алгоритм 2 является переборным. Поэтому основной алгоритмической проблемой является большое число итераций цикла в этом алгоритме. Различные подходы к решению этой проблемы реализованы в моделях морфологического анализа и синтеза словоформ естественного языка.

Основными лингвистическими проблемами являются синонимия и омонимия.

Синонимия — это образование словоформы более чем одним способом. Опишем проблему синонимии следующим образом:

$$\exists lgw_i = (l_i g_i w_i)\ \exists lgw_j = (l_j g_j w_j)\ ((l_i = l_j) \wedge (g_i = g_j) \wedge (w_i \ne w_j)),$$
$$1 \le i \le cLGW, 1 \le j \le cLGW.$$

Проблема синонимии разрешается введением для каждой синонимичной формы отдельного грамматического значения.

Омонимия — это наличие у одной словоформы нескольких грамматических значений:

$$\exists lgw_i = (l_i g_i w_i)\ \exists lgw_j = (l_j g_j w_j)\ ((g_i \neq g_j) \wedge (w_i = w_j)),$$
$$1 \leq i \leq cLGW, 1 \leq j \leq cLGW.$$

Проблема омонимии неразрешима на морфологическом уровне автоматической обработки текстов и разрешается на синтаксическом уровне. Поэтому алгоритмы морфологического анализа могут возвращать не одно, а список возможных решений. Также омонимия накладывает на алгоритмы морфологического анализа дополнительное требование наличия всех решений в этом списке.

5. Модели морфологического анализа и синтеза языка

5.1. Операционное расширение математического описания предметной области

Дополним математическое описание предметной области множествами операций, необходимыми для описания процессов: $H = \{h_1, h_2, ..., h_{cH}\}$ — множество операций над словоформами; $HN = \bigcup_{i=1}^{\max HN} H^i$ — множество последовательностей операций над словоформами.

Одноместная операция $h \in H$ преобразует строку (последовательность символов) s с получением строки s':

$$s' = h(s).$$

Модели морфологического анализа и синтеза языка в зависимости от трактовки формообразования делят на три группы [14]:

1) элементно-комбинаторная, предполагающая, что форма слова состоит из основы и добавляемых к ней аффиксов, каждому из которых соответствует граммема;

2) элементно-операционная, определяющая правила влияния добавляемых к основе аффиксов на другие части словоформы;

3) словесно-парадигматическая, разделяющая все лексемы по типам формообразования; тип формообразования определяет правила образования словоформ.

Модели являются лингвистическими, так как используют особенности языка, которые не определены в математической постановке задачи. Формализуем их на основе математического описания предметной области с использованием математико-алгоритмического подхода.

Элементно-комбинаторная модель. Элементно-комбинаторная модель (Item and Arrangement) морфологического анализа и синтеза предполагает, что слово-

форму можно представить как совокупность основы и одного или нескольких аффиксов. Каждый аффикс соответствует некоторой грамме. Чтобы словоформа имела заданное грамматическое значение, необходимо добавить к основе аффиксы, соответствующие граммемам, составляющим грамматическое значение.

Для описания этой модели и далее будем использовать следующие множества: $B = \{b_1, b_2, ..., b_{cB}\}$ — множество основ (basic form); $M = \{m_1, m_2, ..., m_{cM}\}$ — множество аффиксов.

Предположим для упрощения, что каждой лексеме соответствует единственная основа: $LB \subset L \times B$; $\overline{LB}$.

Порядок аффиксов в словоформе определен: $mn \in M^{cM}$.

Каждому аффиксу соответствуют граммема и операция:

$$MVH \subset M \times V \times H; \overline{MVH}.$$

В модели используется операция, состоящая в добавлении аффикса к словоформе справа:

$$h_i(s) = s + m_i, 1 \le i \le cM,$$

где «+» — операция соединения строк.

Обозначим получение i-й составляющей элемента отношения xyz как $xyz[i]$.

Алгоритм морфологического синтеза в элементно-комбинаторной модели (алгоритм 3) заключается в добавлении аффиксов, соответствующих граммемам грамматического значения, к основе.

Алгоритм 3. Алгоритм морфологического синтеза в элементно-комбинаторной модели

Вход: лексема l, грамматическое значение g.

Выход: словоформа w, соответствующая l и g.

1 $(b) \leftarrow LB[l]$

2 $s = b$

3 **for** $m = mn[i], 1 \le i \le cmn$ **do**

4 $(v, h) \leftarrow MVH[m]$

5 **if** $v \in g$ **then**

6 $s = h(s) = s + m$

7 **end if**

8 **end for**

9 $w = s$

10 **return** w

Введем алгоритмическую функцию endsWith(s, s'), которая возвращает логическое значение «ИСТИНА» в случае, если строка s заканчивается подстрокой s', и логическое значение «ЛОЖЬ» в противном случае.

Введем также операцию «–» отделения аффикса m от словоформы s справа в случае его наличия: $s - m$.

Алгоритм морфологического анализа в этой модели (алгоритм 4) состоит в отделении аффиксов от словоформы и формировании грамматического значения граммемами, соответствующими отделенным аффиксам.

Алгоритм 4. Алгоритм морфологического анализа в элементно-комбинаторной модели

Вход: словоформа w.

Выход: лексема l, грамматическое значение g, соответствующие w.

```
1   s = w
2   g = ∅
3   for m = mn[i], cmn ≥ i ≥ 1 do
4     if endsWith (s, m) then
5       (v, h) ← MVH[m]
6       s = s − m
7       g = g ⋃ {v}
8     end if
9   end for
10  b = s
11  (l) ← LB[b]
12  return l, g
```

В алгоритме 4 и далее предполагается для упрощения, что $\overline{LB}$. В модели используются операции добавления и отделения аффикса. Операция поставлена в соответствие к аффиксу и грамматеме. Грамматическое значение формируется в процессе отделения аффиксов от основы.

Справедливо следующее неравенство: $cW >> cM$.

Поэтому модель позволяет значительно сократить пространство поиска при морфологическом анализе.

5.2. Элементно-операционная модель

Элементно-операционная модель (Item and Process) использует другую особенность морфологии естественных языков. При синтезе словоформ добавляемый аффикс воздействует на основу или аффиксы при определенных условиях. Аффиксу соот-

ветствует одна или несколько операций. Операции не только добавляют аффикс к основе, но и изменяют другие части словоформы. В остальном модель аналогична элементно-комбинаторной модели.

В этой модели используется отношение $MVHN \subset M \times V \times HN$; $\overline{MVHN}$.

Алгоритм морфологического синтеза в элементно-операционной модели (алгоритм 5) аналогичен алгоритму 3. Однако в алгоритме 5 после присоединения аффикса к формируемой словоформе применяется последовательность операций, которая может изменять различные части словоформы.

Алгоритм 5. Алгоритм морфологического синтеза в элементно-операционной модели

Вход: лексема l, грамматическое значение g.

Выход: словоформа w, соответствующая l и g.

```
1   (b) ← LB[l]
2   s = b
3   for m ← mn[i], 1 ≤ i ≤ cmn do
4     (v, hn) ← MVHN[m]
5     if v ∈ g then
6       s = hn(s)
7     end if
8   end for
9   w = s
10  return w
```

Алгоритм 5 отличается от алгоритма 3 строками 4 и 6.

В модели используются более разнообразные операции, чем в элементно-комбинаторной модели. Это может быть не только добавление и удаление аффиксов, но и изменение основы, например контекстная замена. Как и в предыдущей модели, последовательность операций поставлена в соответствие аффиксу.

Элементно-операционная модель также позволяет сократить пространство поиска по тем же причинам, что элементно-комбинаторная модель.

В работе [15] показано, что элементно-комбинаторная и элементно-операционная модели отличаются незначительно с вычислительной точки зрения. В [16] утверждается, что описание в терминах элементно-комбинаторной модели может быть преобразовано в описание в терминологии элементно-операционной модели. Алгоритмы 3 и 5 подтверждают это.

Элементно-операционная модель формализована в теории автоматов двухуровневой моделью морфологического анализа и синтеза К. Коскенниеми [17]. Модель К. Коскенниеми схожа и с элементно-комбинаторной моделью [18].

5.3. Словесно-парадигматическая модель

Элементно-комбинаторная и элементно-операционная модели являются морфемо-ориентированными. В отличие от них словесно-парадигматическая модель (Word and Paradigm) является словоориентированной. В этой модели словоформа рассматривается целиком, а не как последовательность морфем. «Парадигму часто представляют графически как своего рода таблицу с несколькими входами; строки и столбцы этой таблицы содержат названия граммем» [19]. В ячейках таблицы находятся правила, которые позволяют получить форму, соответствующую комбинации граммем [20, 21].

Взаимосвязь множества последовательности операций с другими множествами можно описать отношением $LGHN \subset L \times G \times HN$, $\overline{LGHN}$.

Большое влияние на развитие словесно-парадигматической модели оказали работы [22, 23]. Современное состояние этой модели морфологического анализа и синтеза представлено в [20].

Словесно-парадигматическая модель в таком виде не сокращает пространство поиска. Однако существуют особенности естественных языков, использование которых позволяет уменьшить это пространство.

5.4. Словесно-парадигматическая модель с флективными классами

Если парадигмы некоторых лексем имеют общую структуру и одинаковые значения в ячейках таблицы, то такие лексемы объединяются в флективный класс (см. например в [24, 25]).

Введем дополнительные множества, необходимые для описания этой модели:

$T = \{t_1, t_2, ..., t_{cT}\}$ — флективные классы слов.

$LT \subset L \times T$, $\overline{LT}$.

$TGHN \subset T \times G \times HN$, $\overline{TGHN}$.

Алгоритм морфологического синтеза в словесно-парадигматической модели с флективными классами (алгоритм 6) состоит из следующих шагов. По лексеме определяются основа и флективный класс. По основе и грамматическому значению определяется последовательность операций, которая применяется к основе для получения словоформы.

Алгоритм 6. Алгоритм морфологического синтеза в словесно-парадигматической модели с флективными классами

Вход: лексема l, грамматическое значение g.

Выход: словоформа w, соответствующая l и g.

1 $(b) \leftarrow LB[l]$

2 $(t) \leftarrow LT[l]$
3 $(hn) \leftarrow TGHN[t, g]$
4 $w = hn(b)$
5 **return** w

Операциями в словесно-парадигматической модели могут быть добавления и отделения аффиксов, а также операции элементно-операционной модели [20]. Операции поставлены в соответствие сочетанию флективного класса и грамматического значения.

Введение флективных классов позволяет сократить пространство поиска за счет того, что перебираются не словоформы, а флективные классы, при этом $cW \gg cT$.

5.5. Универсальная модель формообразования

Словесно-парадигматическая модель с флективными классами позволяет описывать флективное образование словоформ. Однако операции, используемые в этой модели, демонстрируются на конкретных примерах и их свойства не оговариваются. В результате исследования и анализа операций образования словоформ была предложена универсальная по отношению к естественным языкам различных групп и семейств модель формообразования. Впервые модель была описана в [26]. В работе [27] приведена ее алгебраическая формализация, а в работе [28] — в терминах теории графов. В терминологии этой модели операции называются преобразованиями, а последовательности операций — цепочками преобразований.

В модели введены понятия обратного преобразования h' и обратной цепочки преобразований hn'.

Для преобразования h существует преобразование $h' \in H$, обратное ему по действию:

$$\forall h_i \exists h'(s = (h'(h_i(s)))), 1 \le i \le cH.$$

Тогда для цепочки преобразований $hn_i \in HN$ существует цепочка $hn' \in HN$ той же длины, обратная по действию:

$$\forall hn_i \exists hn'(s = (hn'(hn_i(s)))), chn_i = chn', 1 \le i \le cHN.$$

Преобразования должны обладать следующими свойствами.

1) однозначность результата: результаты применения преобразования к одной и той же строке конечное число раз должны быть равными;

2) обратимость действия: для каждого преобразования должно существовать обратное по действию преобразование.

В [29] было доказано, что получение любой грамматической формы любого языка (даже не естественного) с морфологией можно представить в виде цепочки преобразований.

Вместо понятия флективного класса в модели используется понятие типа формообразования, так как образование словоформы может включать преобразования основы, а не только добавление флексии.

Модель преобразования лежит в основе алгоритмов морфологического анализа и синтеза словоформ, называемых алгоритмами определения и генерации словоформ соответственно [29].

Будем использовать в алгоритмах следующие отношения:

$HH' \subset H \times H$, $hh_i' = (h'h_i') \in HH'$, $1 \le i \le cHH'$, $\overline{HH}'$ — соответствие преобразования и обратного к нему;

$HNHN' \subset HN \times HN$, $hnhn_i' = (hn_i, hn_i') \in HNHN'$, $1 \le i \le cHNHN'$, $\overline{HNHN}'$ — соответствие цепочки преобразований и обратной к ней.

$LG \subset L \times G$, $\overline{LG}$ — сочетания лексемы и грамматического значения, используемые для представления результата алгоритма определения словоформ.

Алгоритм генерации аналогичен алгоритму 6. Алгоритм определения (алгоритм 7) перебирает все цепочки, применяет их к словоформе w. Если цепочка, обратная данной, применима к словоформе w, то по полученной основе определяется лексема l, а по примененной цепочке — грамматическое значение g. Полученные сочетания лексем l и грамматических значений g заносятся в список возможных решений, что обусловлено омонимией.

Алгоритм 7. Алгоритм определения на основе универсальной модели формообразования

Вход: словоформа w.

Выход: множество LG лексем l и грамматических значений g, соответствующих w.

```
1   LG = ∅
2   for (t_i, g_i, hn') ∈ TGHN, 1 ≤ i ≤ cTGHN do
3     (hn') ← HNHN'[hn']
4     b = hn'(w)
5     if b ∈ B then
6       (l) ← LB[b]
7       (t) ← LT[l]
8       if t == t_i then
9         LG = LG ∪ {(l, g_i)}
```

```
10          end if
11        end if
12      end for
13      return LG
```

Модель формообразования основана на словесно-парадигматической модели, поэтому она также сокращает пространство поиска.

Особенности универсальной модели.

1. Модель разработана на основе анализа существующих моделей, а также методов морфологического анализа и синтеза, в ходе которого были выявлены их преимущества и недостатки [30]. Модель использует исключительно математико-алгоритмический подход. Влияние лингвистического подхода ограничено словесно-парадигматической моделью, на которой основана данная модель.

2. Рассматриваются не флективные классы, а типы формообразования. Тип формообразования объединяет лексемы, имеющие одинаковые структуры грамматических значений и соответствующих им цепочек преобразований (не обязательно связанные с добавлением флексий).

Отличия универсальной модели формообразования от моделей морфологического анализа и синтеза.

1. Введено понятие обратного преобразования (операции) и сформулированы свойства этих преобразований. Преобразования описаны абстрактно без привязки к конкретному математическому формализму. Такое описание позволяет использовать любой математический аппарат (в отличие от привязки к автоматам в [17]) с соблюдением свойств преобразований.

2. Цепочка преобразований соответствует сочетанию типа формообразования t и грамматического значения g. Цепочка не является правилом и не содержит условий применения (в отличие от правил, например, в [21]), кроме условия применимости к словоформе (например, нельзя отделить аффикс, если словоформа им не заканчивается).

3. На основе модели разработаны алгоритмы морфологического анализа и синтеза словоформ. Модель вместе с алгоритмами генерации и программирования представляет собой целостный подход к решению задач морфологического анализа и синтеза.

4. Модель единообразно описывает регулярные слова и слова-исключения, полные и неполные парадигмы, виды преобразований (добавление и отделение аффиксов, контекстно-свободные и контекстно-зависимые замены, редупликацию и др.).

5. Доказана универсальность модели к естественным языкам различных групп и семейств.

6. С помощью модели показано, что формообразование — это алгоритм.

5.6. Одна лексема, несколько основ и универсальная модель формообразования

В математическом описании предполагается, что каждой лексеме соответствует одна основа. Однако в некоторых случаях одной лексеме может соответствовать несколько основ (например, из-за супплетивизма). В этом случае при алгоритмической реализации существуют два альтернативных подхода: либо хранить все основы (Multiple Underlying Form в терминологии [16]), либо хранить только одну основу и получать другие основы с помощью операций (Single Underlying Form [16]). В работах [16, 31] утверждается, что подход с хранением нескольких основ имеет преимущества над подходом с хранением одной основы, среди которых можно выделить следующие:

- описание основ, которые не могут быть преобразованы в одну с помощью текущих теорий фонологии [16];
- более низкая временна́я сложность по сравнению с подходом с хранением одной основы [31].

Несмотря на преимущества подхода с хранением нескольких основ, в предложенной универсальной модели формообразования используется подход с хранением одной основы по следующим причинам.

1. При морфологическом синтезе нет необходимости в дополнительном соответствии между грамматическим значением g и основой b, которую необходимо преобразовывать в словоформу.

2. Основу можно преобразовать в другую основу одним преобразованием. Это преобразование не увеличивает порядок временно́й сложности.

3. Хранение одной основы сокращает мощность множества B.

Таким образом, подход с хранением одной основы не требует адаптации универсальной модели формообразования. За счет этого данный подход получает ряд преимуществ над подходом с хранением нескольких основ.

6. Заключение

На основе вышеизложенных результатов отметить следующее.

1. Модели разрабатываются авторами на основе их знаний о морфологии естественных языков. Ограниченные знания приводят к порождению моделей с ограни-

чениями. Для разработки универсальных моделей необходимы экспертные знания и абстрактное мышление.

2. Обратите внимание, что в статье, посвященной моделям морфологического анализа и синтеза словоформ естественных языков, не приведен ни один пример из естественного языка. Это сделано сознательно для того, чтобы абстрактно представить модели и задачи исключительно с точки зрения математико-алгоритмического подхода. Такой подход позволил наглядно показать сходства и отличия моделей.

3. С лингвистической точки зрения рассмотренные в статье модели различны. Однако с математико-алгоритмической точки зрения — схожи. Этот вывод подтвержден и другими исследователями.

4. В процессе морфологического анализа и синтеза используются операции над словоформами. Операции ставятся в соответствие аффиксам в элементно-комбинаторной и элементно-операционной моделях. Причем в элементно-комбинаторной модели операция добавления аффикса определяется явно. В словесно-парадигматической модели операции различаются в зависимости от флективного класса и грамматического значения.

5. Предложенная на основе словесно-парадигматической модели универсальная модель формообразования не определяет операции (преобразования), но накладывает ограничения на их свойства. В отличие от словесно-парадигматической модели универсальная модель формообразования вместе с алгоритмами генерации и определения на ее основе представляет собой целостный подход к морфологическому анализу и синтезу словоформ естественных языков различных групп и семейств.

6. Все модели позволяют уменьшить пространство поиска при морфологическом анализе. В элементно-комбинаторной и элементно-операционной моделях пространство сокращается за счет представления формообразования как добавления аффиксов с необходимыми граммемами к основе. Словесно-парадигматическая модель сокращается пространство поиска за счет объединения лексем с одинаковым образованием словоформ в флективных классы.

Автор статьи, что проведенное сравнение моделей морфологического анализа и синтеза поможет исследователям выбирать наиболее подходящую модель для решения теоретических и практических задач.

Литература

[1] *Дал У., Дейкстра Э., Хоор К.* Структурное программирование. Пер. с англ. С. Д. Зеленецкого, В. В. Мартынюка, Л. В. Ухова; под ред. Э. З. Любимского и В. В. Мартынюка. — М. : Мир, 1975. Сер. «Математическое обеспечение ЭВМ».

[2] *Hopcroft J. E., Ullman J. D.* Formal Languages and Their Relation to Automata. — Addison-Wesley, 1969.

[3] *Пруцков А. В.* Задачи автоматической обработки текста на естественных языках и возможные математические подходы к их решениям // *Вестник РГРТУ*. 2016. № 55. С. 81–86.

[4] Языкознание. Бол. энцикл. словарь / гл. ред. В. Н. Ярцева. — 2-е изд. — М. : Бол. рос. энцикл., 1998.

[5] *Александров В. В., Макаров Н. П., Шустов А. С.* Автоматизированный анализ и оценка статей коллективных договоров // *Вестник РГРТУ*. 2013. № 45. С. 71–75.

[6] *Гиголаев А. В., Цуканова Н. И.* Анализ семантической близости слов с помощью карт Кохонена // *Вестник РГРТУ*. 2018. № 64. С. 85–91.

[7] *Брумштейн Ю. М., Дюдиков И. А*. Цели, модели и методы поиска аналогов для ИТ-проектов в сфере образования // *Известия ВолгГТУ*. 2016. № 11 (190). С. 76–83.

[8] *Бакиева А. М., Батура Т. В.* Исследование применимости теории риторических структур для автоматической обработки научно-технических текстов // *Cloud of Science*. 2017. Т. 4. № 3. С. 450–462.

[9] *Ломакина Л. С., Суркова А. С.* Теоретические аспекты концептуального анализа и моделирования текстовых структур // *Фундаментальные исследования*. 2015. № 2 (17). С. 3713–3717.

[10] *Поляков Д. В., Митрофанов Н. М., Матвеева А. С.* Метод формализации нечетких коллокаций термов в текстах на основе лингвистических переменных // *Прикаспийский журнал: управление и высокие технологии*. 2015. № 4 (32). С. 167–183.

[11] *Усманов З. Д., Довудов Г. М.* О статистическом портрете таджикского предложения // *Известия Академии наук Республики Таджикистан. Отделение физ.-мат., хим., геол. и техн. наук*. 2017. № 2 (167). С. 42–48.

[12] *Пруцков А. В.* Статический и динамический подходы к проектированию подсистем проверки знаний автоматизированных обучающих систем // *Информационные ресурсы России*. 2006. № 1. С. 8.

[13] *Поляков В. Н.* Программа «Недоросль»: когнитивный подход к исследованию природы семантических связей естественного языка // Тез. 5-й нац. конф. по искусственному интеллекту КИИ-96. — Казань, 1996. С. 98–101.

[14] *Hockett C. F.* Two Models of Grammatical Description // *Word*. 1954. Vol. 10. No. 2–3. P. 210–234.

[15] *Roark B., Sproat R.* Computational Approaches to Morphology and Syntax. — Oxford University Press, 2007.

[16] *Maxwell M.* Two Theories of Morphology, One Implementation. — Mike Maxwell and Summer Institute of Linguistics, Inc., 1998.

[17] *Koskenniemi K.* Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. Publications No. 11. — University of Helsinki, 1983.

[18] *Johannessen J. B.* Is Two-Level Morphology a Morphological Model? // In Proc. of the 7th Nordic Conference of Computational Linguistics (NODALIDA 1989). — 1990. P. 51-59.

[19] *Плунгян В. А.* Введение в грамматическую семантику: грамматические значения и грамматические системы языков мира. — М. : Рос. гос. гуманитар. ун-т, 2011.

[20] *Blevins J. P.* Word and Paradigm Morphology. — 1st ed. — Oxford University Press, 2016.

[21] *Stump G.* Inflectional Morphology. A Theory of Paradigm Structure. — Cambridge University Press, 2001.

[22] *Matthews P. H.* The Inflectional Component of a Word-and-Paradigm Grammar. *Journal of Linguistics*. 1965. Vol. 1. No. 2. P. 139-171.

[23] *Matthews P. H.* Morphology. — Cambridge University Press, 1991.

[24] *Stump G., Finkel R.* A. Morphological Typology: From Word to Paradigm. — Cambridge University Press, 2013.

[25] *Белоногов Г. Г., Богатырев В. И.* Автоматизированные информационные системы. под ред. К. В. Тараканова. — М. : Сов. радио, 1973.

[26] *Пруцков А. В.* Морфологический анализ и синтез текстов посредством преобразований форм слов // *Вестник РГРТА*. 2004. № 15. С. 70–75.

[27] *Пруцков А. В.* Алгебраическое представление модели формообразования естественных языков // *Cloud of Science*. 2014. Т. 1. № 1. С. 88–97.

[28] *Пруцков А. В., Пылькин А. Н.* Информационная система с использованием поиска решений задач генерации и определения в пространстве словоформ // *Вестник РГРТУ*. 2011. № 36. С. 39–43.

[29] *Пруцков А. В.* Генерация и определения форм слов естественных языков на основе их последовательных преобразований // *Вестник РГРТУ*. 2009. № 27. С. 51–58.

[30] *Пруцков А. В., Розанов А. К.* Методы морфологической обработки текстов // *Прикаспийский журнал: управление и высокие технологии*. 2014. № 3 (27). С. 119–133.

[31] *Hausser R.* Three Principled Methods of Automatic Word Form Recognition // Proc. of VEXTAL: Venecia per il Tratamento Automatico delle Lingue. Venice, 1999. P. 91–100.

Автор:

Александр Викторович Пруцков — доктор технических наук, доцент, профессор кафедры «Вычислительная и прикладная математика», Рязанский государственный радиотехнический университет

Cloud of Science. 2018. Т. 5. № 4

Mathematical and Algorithmic Formalization of Models of Morphological Analysis and Synthesis of Word-Forms of Natural Languages

A. V. Prutzkow

Ryazan State Radio Engineering University
Gagarin str., 59/1, Ryazan, Russian Federation, 390005
e-mail: mail@prutzkow.com

Abstract. Models of morphological analysis and synthesis of word-forms are based on the morphology of natural languages, using a linguistic point of view. However, to use the models in natural language processing program systems requires their analysis from mathematical and algorithmic points of view. The purpose of the study is a formal description of models of morphological analysis and synthesis of word-forms and their analysis on various aspects. Aspects are methods of algorithmic realization of morphological analysis and synthesis of word-forms, used operations and its coupling with model objects, approaches to reduce the search space in morphological analysis. Item and Arrangement, Item and Process, Word and Paradigm models are analyzed, as well as our universal model of form-building. Mathematical and algorithmic formalization of models reveal their similarity. We conclude that in all models, except Item and Arrangement, the operations are explicitly unlimited and most often an affix adding or replacing of a part of the stem by the other. The models reduce the search space in morphological analysis.

Keywords: natural language processing, morphological analysis, morphological synthesis, Item and Arrangement model, Item and Process model, Word and Paradigm model.

References

[1] *Dal U., Deykstra E., Khoor K.* (1975) Strukturnoe programmirovanie. Moscow, Mir.
[2] *Hopcroft J. E., Ullman J. D.* (1969) Formal Languages and Their Relation to Automata. Addison-Wesley.
[3] *Prutzkow A. V.* (2016) *Vestnik RGRTU*. 55:81–86.
[4] Yazykoznaniye. Bol. entsikl. slovar' (1998) Moscow, Bol. ros. entsikl.
[5] *Aleksandrov V. V., Makarov N. P., Shustov A. S.* (2013) *Vestnik RGRTU*. 45:71–75.
[6] *Gigolayev A. V., Tsukanova N. I.* (2018) *Vestnik RGRTU*. 64:85–91.
[7] *Brumshteyn Yu. M., Dyudikov I. A.* (2016) *Izvestiya VolgGNU*. **11**(190):76–83.
[8] *Bakiyeva A. M., Batura T. V.* (2017) *Cloud of Science*. **4**(3):450–462.
[9] *Lomakina L. S., Surkova A. S.* (2015) *Fundamental'nyye issledovaniya*. **2**(17):3713–3717.
[10] *Polyakov D. V., Mitrofanov N. M., Matveyeva A. S.* (2015) *Prikaspiyskiy zhurnal: upravleniye i vysokiye tekhnologii*. **4**(32):167–183.
[11] *Usmanov Z. D., Dovudov G. M.* (2017) *Izvestiya Akademii nauk Respubliki Tadzhikistan. Otdeleniye fiz.-mat., khim., geol. i tekhn. nauk*. **2**(167):42–48.
[12] *Prutzkow A. V.* (2006) *Informatsionnyye resursy Rossii*. 1:8.
[13] *Polyakov V. N.* (1996) Programma «Nedorosl'»: kognitivnyy podkhod k issledovaniyu prirody semanticheskikh svyazey yestestvennogo yazyka. Tez. 5-y nats. konf. po iskusstvennomu intellektu KII-96. Kazan'. P. 98–101.
[14] *Hockett C. F.* (1954) *Word*, **10**:386–399.
[15] *Roark B., Sproat R.* (2007) Computational Approaches to Morphology and Syntax. Oxford University Press.

[16] *Maxwell M.* Two Theories of Morphology, One Implementation. Mike Maxwell and Summer Institute of Linguistics, Inc., 1998.
[17] *Koskenniemi K.* (1983) Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production. University of Helsinki, Department of General Linguistics. Publications No. 11.
[18] *Johannessen J. B.* (1989) Is Two-Level Morphology a Morphological Model? In Proc. of NODALIDA-1989, pp. 51–59.
[19] *Plungyan V. A.* (2011) Vvedeniye v grammaticheskuyu semantiku: grammaticheskiye znacheniya i grammaticheskiye sistemy yazykov mira. Moscow, Ros. gos. gumanitar. un-t.
[20] *Blevins J. P.* (2016) Word and Paradigm Morphology, 1st ed. Oxford University Press.
[21] *Stump G.* (2001) Inflectional Morphology. A Theory of Paradigm Structure. Cambridge Univ. Press.
[22] *Matthews P. H.* (1965) *Journal of Linguistics*. **1**(2):139–171.
[23] *Matthews P. H.* (1991) Morphology. Cambridge Univ. Press.
[24] *Stump G., Finkel R. A.* (2013) Morphological Typology: From Word to Paradigm. Cambridge Univ. Press.
[25] *Belonogov G. G., Bogatyrev V. I.* (1973) Avtomatizirovannyye informatsionnyye sistemy. Moscow, Sov. radio.
[26] *Prutzkow A. V.* (2004) *Vestnik RGRTA*. 15:70–75.
[27] *Prutzkow A. V.* (2014) *Cloud of Science.* **1**(1):88–97.
[28] *Prutzkow A. V., Pyl'kin A. N.* (2011) *Vestnik RGRTU*. 36:39–43.
[29] *Prutzkow A. V.* (2009) *Vestnik RGRTU*. 27:51–58.
[30] *Prutzkow A. V., Rozanov A. K.* (2014) *Prikaspiyskiy zhurnal: upravleniye i vysokiye tekhnologii*. 3(27):119–133.
[31] *Hausser R.* (1999) Three Principled Methods of Automatic Word Form Recognition. In Proc. of VEXTAL: Venecia per il Tratamento Automatico delle Lingue. Venice, Italy. Sept., pp. 91–100.