анализа и синтеза имен прилагательных русского языка.

УДК 62.506.2

Е. А. СОЛОВЬЕВА

АНАЛИЗАТОР И КЛАССИФИКАТОР ЛИЧНЫХ ФОРМ ПО НАКЛОНЕНИЮ И ИХ ИССЛЕДОВАНИЕ

Предлагаются модели процессов морфологической классификации личных глагольных форм русского языка по признаку категории наклонения (с учетом и без учета омографии), а также глаголов комплексной парадигмы по признаку категории репрезентации. Исследуются модели классификации в целях их практического использования.

Отметим некоторые особенности постановки задач [1] моделирования процессов классификации с учетом и без учета омографии.

Предъявляя испытуемому пары словоформ и используя его способность работать в режиме нуль-органа, можно сформировать классы эквивалентности (формальные классы или значения) по различным морфологическим признакам. В один и тот же класс испытуемый отнесет словоформы, обладающие одинаковым набором значений (признаков значений) категории, по признаку которой происходит классификация. Реакция испытуемого на любую (даже омографичную) форму будет однозначной, так как каждой словоформе соответствует признак только одного формального класса [2]. При этом задача классификации будет корректно поставленной.

Введем функцию f^{i}_{M} (назовем ее классификатором личных форм по наклонению), описывающую процесс морфологической классификации личных форм из X (X — множество синтетических личных глагольных форм русского языка) по признаку наклонения M с учетом омографии. Множество значений функции f^{i}_{M} состоит из признаков формальных значений формальной категории наклонения [2]. Аналогично можно ввести функцию, описывающую процесс решения любой задачи морфологической классификации (с учетом омографии), например, классификации глаголов комплексной парадигмы (составляющих множество X_k) по признаку R категории репрезентации (функция f^{i}_{R} — комплексный классификатор глагольных форм). Моделирование функций, которые грамотный человек реализует при решении корректно поставленных задач классификации на морфологическом уровне с учетом омографии, безусловно важно в теоретическом, но не менее полезно и в практическом плане. Построение действующих моделей таких функций — необходимая предпосылка создания высококачественных моделей языка.

Рассмотрим функции, описывающие процессы классификации без учета омографии. Исследования процессов классификации словоформ показывают, что, используя в качестве определяемых признаков только грамматические значения (при необходимости и признаки их отсутствия), мы не получим при наличии омографии корректно поставленных задач классификации. В таких случаях не существует функций, которые мы хотим описать (например, функции определения значений наклонения у личных форм), потому что отдельным словоформам (омографичным) нельзя на морфологическом уровне поставить в соответствие однозначный ответ [2]. Функция будет получена, если из области ее определения исключить омографичные словоформы, обладающие различными признаками значений рассматриваемой категории.

Можно также поставить в соответствие омографичным формам только один из характеризующих их признаков значений (или отсутствия значений). Это значит, что реакция испытуемых на омографичных формах будет однозначной, а функции, не определенные на последних, доопределены заранее установленным образом (исходя обычно из потребностей практики). «Насильственное» доопределение функций, в область определения которых не входят омографичные словоформы, позволяет выбрать в качестве объекта исследований функции, описывающие процессы классификации без учета омографии. Так, можно ввести функцию f_M (анализатор личных форм по наклонению), реализуемую при классификации форм из X по признаку M без учета омографии, функцию f_R (комплексный анализатор глагольных форм)—при классификации элементов X_k по признаку R и т. п. Предложенные функции являются приближенными к тем, которые человек реализует, решая рассматриваемые задачи на уровне контекста, когда каждой словоформе может быть поставлен в соответствие однозначный ответ.

Действующие модели не учитывающих омографию функций представляют практический интерес, если их построение основано на результатах статистических обследований текстов, тем более, что такие функции чрезвычайно просты. Они могут иметь и некоторое теоретическое значение. Введение приближенных функций целесообразно, так как число и частота употребления омографичных глагольных форм сравнительно невелики, а частоты употребления омографичных форм, обладающих различных признаками значений рассматриваемой категории, существенно отличаются.

Перейдем к описанию функции $f_M = \langle F_M, X, Y_M \rangle$, где $Y_M = \{M_1 M_2\}$; M_1 — признак изъявительного наклонения, M_2 — повелительного. Формы сослагательного наклонения не входят в X, потому что являются аналитическими.

Грамотный человек легко установит, что f_M (читаю) $= M_1$, f_M (читай) $= M_2$ и т. д. При попытке отыскать f_M (шли) испытуемый окажется в тупике, так как *шли* является омографичной словоформой (изъявительное наклонение от *идти* и повелительное — от *слать*) и ей на морфологическом уровне (т. е. на уровне отдельных словоформ) соответствуют оба значения наклонения. Чтобы в качестве объекта исследования получить функцию, потребуем от человека относить приведенную и все остальные личные формы к изъявительному наклонению, потому что оно встречается во много раз чаще повелительного [3, 4 и др.]. Аналогично мы поступали и при решении задач классификации по признакам категорий времени, числа, лица.

При построении F_M, который представим в виде (2) [1], мы столкнулись со многими затруднениями из-за полного и частичного совпадения форм различных наклонений. Это объясняется тем, что внешними выразителями категории наклонения служат не только формальные признаки словоформ, но и многие другие: интонации, порядок слов, синтаксические связи и т. п., которые могут быть учтены на других уровнях анализа. В результате проведенных исследований категория $S^{\circ}_M = \{S^{\circ}_{M_1}, S^{\circ}_{M_2}\}$, которая не является разбиением [2], заменена разбиением $S^{*}_M = \{S^{*}_{M_1}, S^{*}_{M_2}\}$. Элементам из S_{M_1} модель поставит в соответствие признак M_1, из $S^{*}_{M_2}$ — M_2. $K_M(x, y)$ удобно представить в виде

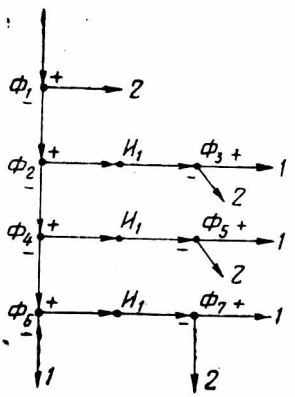
$$K_M(x, y) = [((P(x) \cap P_{M_1} \neq \varnothing) \wedge (P(x) \cap P = \varnothing) \Rightarrow y = M_1) \wedge$$
$$\wedge (P(x) \cap P_{M_1} = \varnothing \Rightarrow y = M_2)],$$

$P_{M_1} = \{\tilde{y}, \tilde{ю}, \tilde{м}, \tilde{шь}, \tilde{ите}, \tilde{ете}, \tilde{т}, \tilde{л}, \tilde{ла}, \tilde{ло}, \tilde{ли}, \tilde{з}, \tilde{с}, \tilde{б}, \tilde{п}, \tilde{к}, \tilde{р}, \tilde{х}, \tilde{г}, \tilde{есть}, \tilde{суть}\}$, $P = \{ляг\}$. Знак „$\sim$“ означает возможность наличия постфикса *ся* (*сь*) [2].

Если из P_{M_1} исключить *ите*, то получим функцию f^{n}_{M}, модель которой целесообразно использовать при действии на разговорных текстах. Интересно принимать во внимание также относительную употребляемость форм изъявительного и повелительного

наклонений различных глаголов, но в таком случае не получим простых моделей, которые полезны для практики, поэтому данный вопрос здесь не рассматривается.

Функция f_M реализована в виде алгоритма A_M, блок классификации которого приведен на рис. 1. Благодаря введению общей схемы алгоритмов классификации глагольных форм [5], для задания алгоритмов достаточно описать лишь их блоки классификации. Для задания блока классификации достаточно указать множества признаков, соответствующих элементарным распознавателям, входящим в блок [1]: $И_j (j=1, 2)$ отбрасывает j последних букв слова. $P_{Ф_0}=\{$ся, сь$\}$, $P_{Ф_1}=\{$й, ка$\}$, $P_{Ф_2}=\{$и, о$\}$, $P_{Ф_3}=\{$л$\}$, $P_{Ф_4}=\{$ь$\}$, $P_{Ф_5}=\{$ш, ест, сут$\}$, $P_{Ф_6}=\{$е$\}$, $P_{Ф_7}=\{$ет, ит$\}$. Алгоритм A_M^n, реализующий f_M^n, получим, если из $P_{Ф_7}$ исключим *ит*.



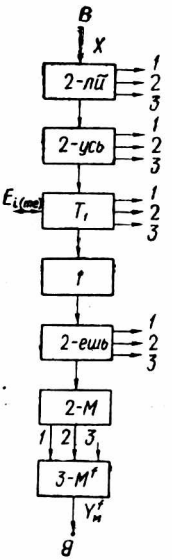


Рис. 1. Граф-схема блока классификации анализатора и классификатора (блок 2-*М*) по наклонению.

Рис. 2. Блок-схема классификатора по наклонению.

Классификатор личных форм по наклонению f_M^i удобно представить в виде блок-схемы алгоритма A_M^i (рис. 2). Блок классификации 2-*М* описан выше (рис. 1), уточняющие блоки предложены в работе [6], а правила объединения блоков с целью построить модификацию модели — в работе [1]. Разработке алгоритма предшествовали длительные и трудоемкие исследования процессов классификации омографичных форм, результаты которых приведены в [6].

На основании этих исследований [6] получены точные модели A_T^i, A_N^i и A_L^i классификации личных форм по признакам категорий времени, числа и лица соответственно с учетом омографии. Модели A_M^i и A_T^i на уровне блок-схемы отличаются только бло-

ками классификации (естественно, необходима перенумерация выходов из блоков [6]). В алгоритм A_N^f входят блоки: 2-ли, 1, H_{eM}^l, 2-N (классификации) и выходной. Блоки классификации алгоритмов A_T^f и A_N^f в составе соответствующих приближенных алгоритмов описаны в работе [7]. В модель A_L^f входят уточняющие блоки 2-ли и 2-усь. Приближенные модели функций, реализуемых при классификации личных форм по признакам времени, числа и лица без учета омографии, приняты в Республиканский фонд алгоритмов и программ.

Полученные модели реализованы на ЭЦВМ «Урал-14Д» (с помощью транслятора АЛГОЛ-ЦЭМИ) и детально исследованы. Приведем некоторые результаты исследований моделей классификации по признаку M.

Алгоритм A_M^f, благодаря методам построения, безошибочно функционирует на любой личной форме глаголов из «Орфографического словаря русского языка» на 104 тыс. слов (11-е издание). При действии A_M^f на различных текстах ошибок также не было обнаружено. Вероятность достоверного предсказания модели A_M, определенная на данных частотного словаря [4] с точностью $\varepsilon=0{,}002$ и вероятностью $p(t)=0{,}95$, составляет $0{,}991\pm0{,}002$, модели A_M^n — $0{,}994\pm0{,}002$ (при составлении словаря [4] использовались и разговорные тексты). Алгоритмы A_M и A_M^n безошибочно действовали на различных выборках личных форм [1]. Несмотря на то что уровень вероятностей достоверного предсказания моделей получился высоким, он не дает нам полного представления о ценности моделей, об их функционировании на различных классах словоформ.

В целях более полной характеристики моделей их исследовали на основании разработанных критериев [2]. Эти критерии отражают специфику текста. Мы выделяли большие группы текстов: литературные и математические, газетные и технические, разговорные и неразговорные и т. п., которые отличаются многими параметрами. Можно рассматривать и более тонкие различия: жанры, стиль авторов и т. п. К сожалению, этот вопрос еще недостаточно изучен в статистической лингвистике и к настоящему времени не имеется точной статистической картины различных текстов.

В интересующем нас аспекте вопрос о частоте употребления слов и форм наиболее полно изучен в работе Э. А. Штейнфельдт [4] для литературных текстов. Математические и технические тексты менее разнообразны и менее богаты употреблением различных слов и форм. Обилие терминов в специальной литературе не повлияет значительно на действие моделей, так как новые термины образуются по правилам образования слов продуктивных групп, а исключения обычно относятся к непродуктивным. Различия в употреблении форм (особенно в неразговорных текстах) для различных групп текстов часто сохраняют некоторую общую

тенденцию. Введенные критерии [2] мы будем определять для наиболее сложного случая — литературных текстов.

При вычислении элементов вероятностных матриц [2] могут возникнуть затруднения, если при классификации словоформ из некоторого класса входных сигналов модели ошибаются очень редко, а частоты употребления таких словоформ сравнительно невелики. Предложенные модели действуют с очень высокой вероятностью достоверного предсказания, а их неточные ответы возможны лишь на незначительной части глаголов повелительного наклонения, причем последние обычно малоупотребительны. Так, в выборках $n_1=1000$ и $n_2=2200$ личных форм из газетных текстов встретилось соответственно лишь два («поезжай-ка», «убирайтесь») и три («выполняйте», «плачь», «ропщи») глагола с признаком M_2, но они классифицируются точно даже приближенными моделями.

На этом примере видно, какое колоссальное количество словоформ и текстов необходимо рассмотреть для накопления сведений о функционировании построенных моделей классификации. Обычный путь отыскания матриц в данном случае нецелесообразен, поэтому предложим более экономный, основанный на использовании частотного словаря [4] (ЧС). Практический подсчет вероятностей, значения которых не очевидны из грамматических соображений, будем производить с помощью данных ЧС на основании методов статистических вычислений. При этом допускаем, что распределение ошибок в тексте и ЧС равновероятно.

Действие модели A_M описывают матрицы P'_M и V'_M [2]:

$$P'_M=\begin{pmatrix}1 & 0\\ 0{,}2\pm 0{,}02 & 0{,}8\pm 0{,}02\end{pmatrix} \text{ и } V'_M=\begin{pmatrix}0{,}95 & 0{,}05\\ 1 & 0{,}8\pm 0{,}02\end{pmatrix}.$$

V'_M показывает, какую роль играют вероятности p_{11} и p_{22} для характеристики A_M. Глаголы изъявительного наклонения, поступающие на вход A_M с вероятностью 0,95, классифицируются безошибочно ($p_{11}=1$); неточные ответы модели могут быть лишь в глаголах повелительного наклонения, встречающихся очень редко (с вероятностью 0,05) даже в литературных текстах ($p_{22}=0{,}8\pm\pm 0{,}02$).

Если считать, что модель A_M, разбивающая множество X на подмножества $S^*_{M_1}$ и $S^*_{M_2}$, разбивает X на три класса S'_{M_1}, S'_{M_2} и S'_{M_3} [2], где $S'_{M_3}=\varnothing$, то функционирование A_M подробно и наглядно опишут матрицы

$$P_M=\begin{pmatrix}1 & 0 & 0\\ 0{,}16\pm 0{,}02 & 0{,}84\pm 0{,}02 & 0\\ 1 & 0 & 0\end{pmatrix},$$

$$V_M=\begin{pmatrix}0{,}935 & 0{,}045 & 0{,}02\\ 1 & 0{,}84\pm 0{,}02 & 0\end{pmatrix}.$$

Процесс расформирования класса $S^{f}_{M_3}$ [2] описывается

$$P^3_M = \begin{pmatrix} p^3_{11} & p^3_{12} \\ p^3_{21} & p^3_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix}$$

и

$$V^3_M = \begin{pmatrix} p^3_1 & p^3_2 \\ p^3_{11} & p^3_{22} \end{pmatrix} = \begin{pmatrix} 0{,}76 & 0{,}24 \\ 1 & 0 \end{pmatrix}.$$

Все матрицы, характеризующие A^f_M и другие модели, которые учитывают омографию, совпадают с единичными.

Проведенные исследования не только позволяют оценить модели и наглядно проследить их функционирование, но и сделать важные выводы о возможности усовершенствования моделей. В алгоритм A_M, например, была введена небольшая поправка и получен алгоритм $A^{\text{п}}_M$, вероятностные характеристики которого еще лучше, чем у A_M:

$$P_{M^{\text{п}}} = \begin{pmatrix} >0{,}997 < 0{,}003 & 0 \\ <0{,}025 > 0{,}975 & 0 \\ 0{,}5 \pm 0{,}03 \;\; 0{,}5 \pm 0{,}03 & 0 \end{pmatrix}.$$

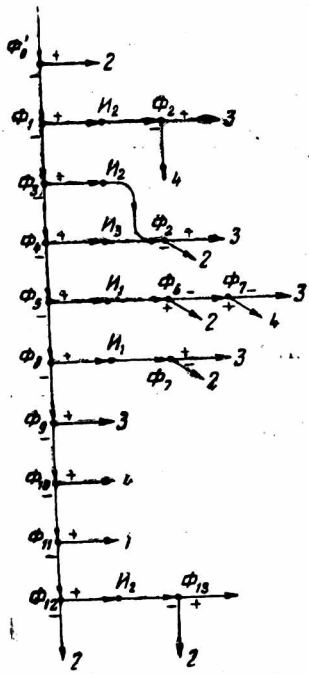


Рис. 3. Граф-схема блока классификации комплексного анализатора и классификатора.

В результате исследований сделаны важные выводы о практическом действии моделей, получены их сравнительные характеристики, подтверждена универсальность и высокая точность.

Для выделения личных глагольных форм из всего множества словоформ целесообразно решить задачу классификации глаголов комплексной парадигмы по признаку категории репрезентации. Эта задача является составной частью задачи классификации словоформ по признаку части речи.

Термин «категория репрезентации» введен А. И. Смирницким для английского языка, а затем применен для русского в работах [8, 9].

Обозначим грамматическую категорию репрезентации через $S^{0}_R = \{S^{0}_{R_1}, \ldots, S^{0}_{R_4}\}$ где $S^{0}_{R_1}$ обозначает множество глаголов в инфинитиве, $S^{0}_{R_2}$ — в личных формах, $S^{0}_{R_3}$ — множество причастий, $S^{0}_{R_4}$ — деепричастий. При построении A_R, реализующего f_R, категорию S_R (которая не является разбиением, так как образующие ее значения, например, $S^{0}_{R_1}$ и $S^{0}_{R_2}$ пересекаются) заменим разбиением $S^{*}_R = \{S^{*}_{R_1}, \ldots, S^{*}_{R_4}\}$.

Модель A_R можно представить в виде общей схемы [5], в которой используется блок полной нормализации [1]. Множеством входных сигналов является множество X_k, выходных — Y, где

сигнал $y_j \in Y$ $(j=\overline{1,4})$ соответствует принадлежности входного сигнала значению S_{R_j}. Блок классификации 2-R показан на рис. 3. Распознаватель Φ_0 проверяет конец слова на-*ка* Φ_1 — на *ая*. Φ_2 — на *ущ, ющ, ащ, ящ, н, т, ем, им, ом,* или *ш* (с предшествующей согласной); Φ_3 — на *ее, ей, ем, ею, ие, ий, им, их, ое, ой, ом, ою, ые, ый, ым, ых* или *ую*; Φ_4 — на *его, ему, ими, ого, ому* или *ыми*; Φ_5 — на *а*; Φ_6 — на *л*; Φ_7 — на *н, т* или *м*; Φ_8 — на *о*; Φ_9 — на *н, ы, от, рт, ыт, нут* или *м*; Φ_{10} — на *в, я* или *ши* (с предшествующей согласной); Φ_{11} — на *ть, чь*; Φ_{12} — на *ти*; Φ_{13} — на *с, з, й, д*. В программе проверка на *ш* или *ши* (с предшествующей согласной) заменена более простой для ЭЦВМ проверкой на *ш* или *ши* (с предшествующей гласной) с соответствующей заменой номеров выходов.

Для построения модели A_R^f, реализующей функцию f_R^f (с учетом омографии), были проведены экспериментальные исследования процессов классификации глаголов из X_k по признаку R на уровне слов и псевдослов и тщательный формальный анализ таких глаголов; выявлены и разрешены все случаи омографии. Категория S_R^0 заменена формальной категорией $S_R^f = \{S_{R_1}^f, \ldots, S_{R_4}^f\}$, где $S_{R_1}^f = S_{R_1}^0 \setminus S_{R_2}^0$, $S_{R_2}^f = S_{R_2}^0 \setminus (S_{R_1}^0 \cup S_{R_3}^0)$, $S_{R_3}^f = S_{R_3}^0 \setminus S_{R_2}^0$, $S_{R_4}^f = S_{R_4}^0$, $S_{R_5}^f = S_{R_1}^0 \cap S_{R_2}^0$, $S_{R_6}^f = S_{R_2}^0 \cap S_{R_3}^0$. Класс $S_{R_5}^f$ содержит омографичные глагольные формы, оканчивающиеся на *ть* и *ти*; $S_{R_6}^f$ — на *ем, им* и *т*.

Приведем алгоритмическую модель A_R^f, разработанную с помощью статистического метода словарей [6] на основании данных ЧС. Выходной сигнал y_j $(j=\overline{1,6})$ модели A_R^f соответствует принадлежности входного слова к формальному классу $S_{R_j}^f$. A_R^f можно получить, добавив перед моделью A_R уточняющий блок. Структура последнего очень проста: он состоит из распознавателей $\Phi_1^c \div \Phi_5^c$ [6], соединенных последовательно, причем отрицательный выход Φ_1^c соединен с выходом Φ_2^c и т. п., отрицательный выход последнего распознавателя Φ_5^c подсоединяется к алгоритму A_R. Положительные выходы распознавателей отмечены цифрами *5, 2, 3, 4* и *6* соответственно. $P_{\Phi_1^c} = \{$есть$\}$, $P_{\Phi_2^c} = \{$ плачь, пусти, выпусти, ответь, спрячь $\}$, $P_{\Phi_3^c} = \{$занят, принят, взят, одет, разбит, поднят $\}$, $P_{\Phi_4^c} = \{$ поднимая $\}$, $P_{\Phi_5^c} = \{$ любим $\}$.

Алгоритм A_R^f безошибочно функционирует на любой из синтетических форм глаголов, входящих в словарь [4]. Вероятностные оценки модели A_R также достаточно высоки. Например, на выборке $n_R = 600$ глагольных форм из литературного текста алгоритм не сделал ни одной ошибки.

Комплексные анализатор и классификатор глагольных форм реализованы на ЭЦВМ, A_R принят в Республиканский фонд алгоритмов и программ [10]. Предложенные модели вошли в комплекс

алгоритмов и программ для подсистемы морфологического анализа автоматизированной системы лингвистической обработки документов.

СПИСОК ЛИТЕРАТУРЫ

1. Соловьева Е. А. Моделирование процессов морфологической классификации с учетом омографии. — См. статью в настоящем сборнике.
2. Соловьева Е. А. К вопросу о построении общего алгоритма классификации глагольных форм русского языка. — В кн.: Проблемы бионики. Вып. 15. Харьков, 1975, с. 143—149.
3. Йоссельсон Г. Г. Подсчет слов и частотный анализ грамматических категорий русского литературного языка. — В кн.: Автоматизация в лингвистике. — М.-Л., «Наука», 1966, с. 105—131.
4. Штейнфельдт Э. А. Частотный словарь современного русского литературного языка. Таллин, 1973. 316 с.
5. Соловьева Е. А. Автоматический морфологический анализ суженной парадигмы глагола. — В кн.: Проблемы бионики. Вып. 12. Харьков, 1974, с. 139—142.
6. Соловьева Е. А. Исследование процессов классификации омографичных глагольных форм. — В кн.: Проблемы бионики. Вып. 16. Харьков, 1976, с. 104—114.
7. Шабанов-Кушнаренко Ю. П., Соловьева Е. А. Бионические аспекты моделирования речевого поведения человека. — В кн.: Проблемы бионики. Вып. 13. Харьков, 1974. с. 59—66.
8. Волоцкая З. М., Молошная Т. Н., Николаева Т. М. Опыт описания русского языка в его письменной форме. М., «Наука», 1964. 186 с.
9. Пиотровская А. А. Машинная морфология русского глагола. — В кн. Статистика речи и автоматический анализ текста. Л., «Наука», 1973, с. 260—277.
10. Мурашко А. Г., Соловьева Е. А. Алгоритм автоматического определения значений категории репрезентации. — РФАП АН УССР, Справка №71 Киев, 1974.