13. Борн М. Физика в жизни моего поколения. М., ИЛ, 1963. 535 с.
14. Балонов Л. Я. Последовательные образы. Л., «Наука», 1971. 214 с.
15. Бугай Ю. П. Об особенностях элементарных и неэлементарных фор отражения, существенных для моделирования нервной системы. Сообще ние I (см. ст. в настоящем сборнике).
16. Ярбус А. Л. Роль движения глаз в процессе зрения. М., «Наука» 1965. 166 с.

УДК 62.506.2

Ю. П. ШАБАНОВ-КУШНАРЕНКО, д-р техн. наук
Е. А. СОЛОВЬЕВА, инж.

К ВОПРОСУ ОБ АВТОМАТИЧЕСКОМ МОРФОЛОГИЧЕСКОМ АНАЛИЗЕ ПРИЧАСТИЙ РУССКОГО ЯЗЫКА

Математическое моделирование на ЭЦВМ способности человека решать задачи морфологического анализа позволяет автоматически получать морфологическую информацию о словоформе. При этом наибольшую практическую и теоретическую ценность представляет изучение словесного поведения идеально грамотного человека, которое мы и анализируем.

В настоящей работе исследуется классификация причастий по признаку формы (полной или краткой), залога и времени. Причастия входят в систему глагольных образований. Поэтому необходимо научить ЭЦВМ автоматически отличать их от остальных форм глагола. Этот вопрос рассматривается здесь в более общем виде как задача определения, к какому из множеств (глаголов, причастий или деепричастий) принадлежит любая глагольная форма.

Изложим некоторые соображения относительно выбора задач для моделирования.

Причастие и деепричастие являются атрибутивными формами глагола, обозначающими действие как признак, свойство предмета или лица или как признак, характеризующий другое действие [1]. Причастие совмещает в себе признаки глагола и прилагательного, но теснее связано с глаголом. Это проявляется прежде всего в общности основы и ее лексического значения, а также в наличии у причастия основных глагольных категорий, например залога и времени [2, 3]. Деепричастие, обозначающее в качестве наречного признака тот процесс, который назван инфинитивом, относится вместе с инфинитивом к морфологическому разряду неизменяемых слов, но при этом также ближе к глаголу и является его формой. Следовательно, в данной работе моделируются задачи определения глагольных признаков причастий и различения основных множеств глагольных форм.

Введем обозначения, которые позволят описать функционирование моделей. Множество глагольных форм русского языка обозначим через $X = \{x_1, \ldots, x_n, \ldots\}$, глаголов — $X^V = \{x_1^V, \ldots$

$\ldots, x_k^V, \ldots\}$, причастий — $X^P = \{x_1^P, \ldots, x_m^P, \ldots\}$, деепричастий — $X^A = \{x_1^A, \ldots, x_l^A, \ldots\}$. Число n, k, m, l элементов этих множеств не ограничено ввиду развития языка и возможности появления в нем новых словоформ. Для обозначения произвольного элемента

$$x_i \in X, \tag{1}$$

$$x_i^V \in X^V, \tag{2}$$

$$x_i^P \in X^P, \tag{3}$$

$$x_i^A \in X^A, \tag{4}$$

например, глагольной формы *читаю*, будем применять запись $x_i = [читаю]$ или $x_i^V = [читаю]$, так как известно, что $([читаю] \in X) \wedge ([читаю] \in X^V)$. Следует отметить, что x_i является синтетической (одно слово) формой глагола. В данной работе примеры глагольных форм приводятся из словаря [4].

Множества X^V, X^P, X^A являются подмножествами множества X ($X^V \subset X$, $X^P \subset X$, $X^A \subset X$), причем

$$X = X^V \cup X^P \cup X^A. \tag{5}$$

Эти подмножества могут в принципе пересекаться, например:

$$X^V \cup X^P \neq \varnothing$$

Покажем это. Так, $x_i = [изменяем]$ относится одновременно к множествам глаголов и причастий, т. е. выполняется условие $x_i \in X^V \cap X^P$. Следовательно, (5) верно.

Множество $Y^P = \{y_1^P, \ldots, y_r^P\}$ объединяет r признаков (выбрано $r = 6$) причастий, приведенных в таблице Y_j^P.

Форма	Время	Залог	
		Действительный	Страдательный
Полная	Настоящее	y_1^P	y_2^P
	Прошедшее	y_3^P	y_4^P
Краткая	Настоящее	—	y_5^P
	Прошедшее	—	y_6^P

Элемент y_j ($j = \overline{1,3}$) множества $Y = \{y_1, y_2, y_3\}$ означает принадлежность к одному из трех подмножеств: X^V (при $j = 1$), X^P ($j = 2$), X^A ($j = 3$).

Преобразователь словесной информации F^P, ставящий в соответствие произвольному элементу $x_i^P \in X^P$ некоторый элемент $y_j^P \in Y^P$ (X^P — множество входных слов F^P; Y^P — выходных), является математической моделью способности человека опреде-

лять форму, залог и время причастий русского языка. Эта модель получена в виде граф-схемы алгоритма Γ^P, которую можно представить как последовательное соединение трех составных блоков — нормализации, классификации и выходного. Аналогичный вид (с точностью до обозначений) имеет алгоритм морфологического анализа A_k (рис. 1 [5]). Блок нормализации алгоритма идентичен блоку *1* (рис. 2, *а*, [5]). Выходной блок Γ^P подобен блоку *3* (рис. 2*б*, [5]), но так как определенное количество выходных сигналов алгоритма (для $\Gamma^P - r = 6$) требует такого же числа операторов в выходном блоке и соответственно входов в этот блок, для выходного блока Γ^P окажется шесть операторов и входов.

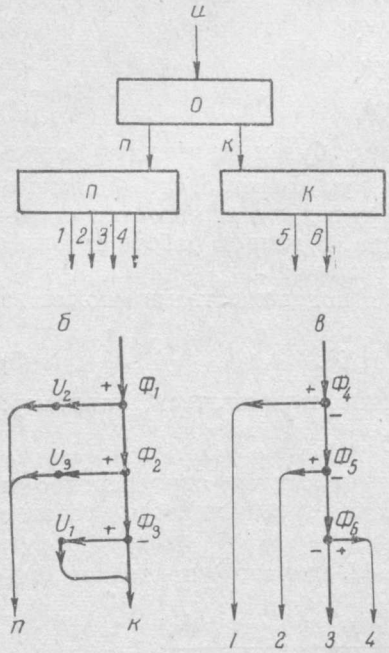

Рис. 1.

Блок классификации, показанный на рис. 1, *а*, удобно подразделить на три блока: *О* — выделения основы, *П* — классификации полных причастий, *К* — классификации кратких. Выделяя основу причастия, блок *О* направляет ее на выход *П*, если причастие стоит в полной форме, и на выход *К* — если в краткой. Блок *О* (рис. 1, *б*) состоит из операторов $U_1 - U_3$ и распознавателей $\Phi_1 - \Phi_3$. Операторы U_1, U_2, U_3 отбрасывают соответственно одну, две или три последние буквы слова. Распознаватели проверяют конец слова на совпадение с одним из буквосочетаний *ая*, *ее*, *ей*, *ем*, *ею*, *це*, *ий*, *им*, *их*, *ое*, *ой*, *ом*, *ою*, *ые*, *ый*, *ым*, *ых*, *ую* (Φ_1); *его*, *ему*, *ими*, *ого*, *ому*, *ыми* (Φ_2) или с одной из букв *а*, *о*, *ы* (Φ_3). В состав блока *П* (рис. 1, *в*) входят только распознаватели ($\Phi_4 \div \Phi_6$), поэтому он не изменяет сигнал, поступающий на его вход. Φ_4 проверяет конец слова на *ущ*, *ющ*, *ащ* или *ящ*, Φ_5 — на *н* или *т*, Φ_6 — на *ем*, *им* или *ом*. Блок *К* состоит из одного распознавателя Φ_5, положительный выход которого отмечен цифрой *6*, а отрицательный — цифрой *5*.

При отладке алгоритма Γ^P на большом массиве слов (3) не было обнаружено ни одного x_i^P, неточно проанализированного. Эти результаты, а также данные исследования множества X^P и правил его образования позволяют заключить, что погрешность алгоритма $\Gamma^P - \varepsilon^P = 0$ (видимо, теоретически следует считать $\varepsilon^P \approx 0$).

Рассмотрим преобразователь F, который любому элементу x_i входного множества X ставит в соответствие элемент y_j выходного множества Y. Этот преобразователь, моделирующий поведение человека при классификации глагольных форм на три разряда (глаголы, причастия или деепричастия), получен в виде алгоритма Г, аналогичного по своей структуре Γ^P. Подавая на вход Г слово (1), на его выходе получаем признак y_1, y_2 или y_3 в зависимости от того, выполнение какого из условий (2), (3) или (4) он установит. Граф-схема упрощенного блока классификации алгоритма Г приведена на рис. 2. Она состоит из распознавателей $\Phi_1 - \Phi_3$ и операторов $U_1 - U_3$. Распознаватель Φ_1 проверяет конец слова на *ая*; Φ_5 — на *а* или *о*; Φ_6 — на *л*; Φ_8 — на *н*, *ы*, *от*, *рт*, *ыт*, или *нут*; Φ_9 — на *в*, *я* или *ши* (*с предшествующей согласной*). Выполняются условия

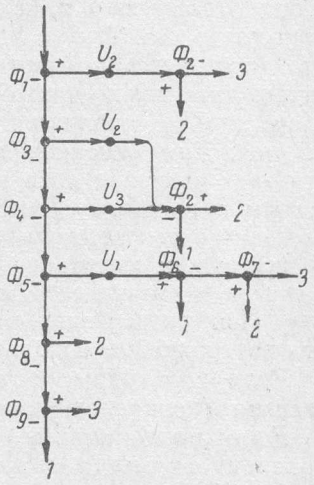

Рис. 2.

$$[\Phi_4]_\Gamma = [\Phi_2]_{\Gamma^P},\ [\Phi_7]_\Gamma = [\Phi_5]_{\Gamma^P},$$
$$[U_1]_\Gamma = [U_1]_{\Gamma^P},\ [U_2]_\Gamma = [U_2]_{\Gamma^P},$$
$$[U_3]_\Gamma = [U_3]_{\Gamma^P},$$

означающие идентичность функционирования соответствующих блоков алгоритмов Г и Γ^P, а также условие

$$[\Phi_3]_\Gamma = [\Phi_1]_{\Gamma^P} - [\Phi_1]_\Gamma,$$

из которого следует, что Φ_3 в алгоритме Г проверяет конец слова на те же буквосочетания, что и Φ_1 в алгоритме Γ^P, за исключением буквосочетания *ая*. Φ_2 проверяет конец слова на *ущ*, *ющ*, *ащ*, *ящ*, *н*, *т*, *ем*, *им* или *ом*, т. е.

$$[\Phi_2]_\Gamma = [\Phi_4]_{\Gamma^P} + [\Phi_5]_{\Gamma^P} + [\Phi_6]_{\Gamma^P}.$$

Ввиду сложности точного алгоритма Г в данной работе приведен упрощенный блок классификации (рис. 2), не всегда дающий верное решение. Число глагольных форм, анализируемых ошибочно, настолько мало по сравнению со всем множеством X, что погрешность алгоритма ε будет незначительно отличаться от нуля. Эту погрешность можно уменьшить (при условии, что множество X конечно) путем введения перед блоком классификации дополнительного блока, который позволит получить более точный (но усложненный) алгоритм Γ_1.

Не описывая дополнительный блок, отметим наиболее важные случаи, которые будут им рассматриваться.

Подавляющее большинство глагольных форм можно классифицировать однозначно. Возможен и неоднозначный ответ. Например, для $x_i = [\textit{изменяем}] \in X \cap X^P$ правильным будет ответ

$y_4 = y_1 \wedge y_2$, а алгоритм F выдаст сигнал y_1. Число таких случаев незначительно, потому что краткие страдательные причастия настоящего времени на *ем*, *им* в современном русском языке почти не употребляются [6], но при необходимости их можно учитывать.

Рассмотрим $x_i = [подшéй]$ и $x_q = [пáдшей]$. Алгоритм F, проверив концы слов на совпадение с *ей*, а затем, отбросив *ей*,— с *дш*, отнесет x_i и x_q к причастиям, в то время как x_l является глаголом. Такую ошибку можно устранить, если заметить, что ударение в причастиях не падает на флексии. Избавиться от ошибки можно также путем различения приставки и корня словоформ. Деепричастия, как и причастия, могут оканчиваться на *тая*. Для устранения неточности в этом случае можно ввести знак ударения либо воспользоваться другими приемами. Алгоритм F входным сигналам $x_l = [пропащим]$ и $x_q = [протащим]$ поставит в соответствие один и тот же выходной сигнал y_2, в то время как x_q должен соответствовать y_1. Такие x_q малочисленны и могут классифицироваться точно дополнительным блоком. При разрешении подобных случаев мы стремимся, когда это возможно, не использовать словарь исключений.

Если необходимо автоматически различать четыре формы глагола (неопределенную, личную, причастие и деепричастие), то F можно дополнить алгоритмом классификации глаголов по признаку формы (неопределенной или личной), предложенным в работе [7].

Рассмотренные алгоритмы реализованы на ЭЦВМ и могут применяться на практике.

ЛИТЕРАТУРА

1. Грамматика русского языка. Т. I. М., Изд-во АН СССР, 1960. 719 с.
2. Грамматика современного русского литературного языка. М., «Наука», 1970. 767 с.
3. Гужва Ф. К. Современный русский литературный язык. Киев, «Радянська школа», 1967. 223 с.
4. Орфографический словарь русского языка, изд. 11-е. М., «Сов. энциклопедия», 1971. 520 с.
5. Соловьева Е. А. Автоматический морфологический анализ суженной парадигмы глагола (см. ст. в настоящем сборнике).
6. Земский А. М., Крючков С. Е., Светлаев М. В. Русский язык. ч. 1-я. М., «Просвещение», 1971. 287 с.
7. Бондаренко М. Ф., Соловьева Е. А. Методы решения задач морфологической и субморфологической классификации.— В сб.: Проблемы бионики. Вып. 10. Харьков, 1973, с. 145—150.