УДК 62.506.2

Е. А. СОЛОВЬЕВА, инж.

АВТОМАТИЧЕСКИЙ МОРФОЛОГИЧЕСКИЙ АНАЛИЗ СУЖЕННОЙ ПАРАДИГМЫ ГЛАГОЛА

В настоящей работе исследуется и моделируется поведение испытуемого при классификации суженной парадигмы глагола по признакам времени (задача 1), числа (задача 2), лица (задача 3) и рода (задача 4). Человек способен решать вышеназванные задачи морфологической классификации глаголов русского языка. Опишем математические модели такой способности идеально грамотного человека.

Суженную парадигму глагола необходимо автоматически анализировать потому, что к ней относится большинство глаголов, встречающихся в различных текстах (технических, литературных, газетных) и в разговорной речи, а алгоритмы морфологической классификации этих глаголов (и составление алгоритмов) сравнительно просты. В суженную парадигму глагола входят синтетические (простые) формы изъявительного наклонения, выражающие основные категориальные значения слов, т. е. формы настоящего, настоящего — будущего (для глаголов совершенного вида) и прошедшего времени [1].

Математическая модель решения каждой из вышеназванных задач анализа суженной парадигмы глагола в виде алгоритма $A_к$ ($к = \overline{1,4}$) показана на рис. 1. Множество $X = \{x_1, \ldots, x_n\}$ входных сигналов алгоритма $A_к$ составляют формы всех глаголов русского языка (для определенности взятых из словаря [2] на 104 тыс. слов), входящих в суженную парадигму. Множество выходных сигналов $Y_к = \{y_{к1}, \ldots, y_{кm}\}$ — это набор признаков, определяемых в конкретной задаче классификации (K — номер задачи). Подавая на вход B алгоритма $A_к$ какую-либо глагольную форму $x_i \in X$, на его выходе g получаем признак $y_{кj} \in Y_к$, соответствующий входному слову x_i.

Рассмотрим более подробно выходные множества $Y_1 - Y_4$, содержательное описание которых приведено в таблице $y_{кj}$.

k	j			
	1	2	3	4
1	Непрошедшее время	Прошедшее время	—	—
2	Единственное число	Множественное число	—	—
3	1-е лицо	2-е лицо	3-е лицо	Нет лица
4	Мужской род	Женский род	Средний род	Нет рода

Сигналы $Y_{34} \in Y_3$ или $Y_{44} \in Y_4$ означают соответственно отсутствие признаков лица или рода у входного слова. Эти сигналы необходимы потому, что частные грамматические значения лица или рода имеет лишь часть глаголов суженной парадигмы. Точнее, все множество X можно разбить на два непересекающихся подмножества таким образом, что глагольные формы, входящие в одно из них, будут иметь значения лица (но не иметь рода), а входящие в другое — значения рода (не имея лица).

В задаче 1 будем различать два выходных сигнала: y_{11} (непрошедшее время) и y_{12} (прошедшее время). Формы непрошедшего времени объединяют глаголы настоящего (несовершенный вид) и будущего (совершенный вид) времени, временные значения которых обусловлены видом глагола. Такое объединение необходимо, потому что категорию вида нецелесообразно классифицировать автоматически на морфологическом уровне (это удастся лишь на смысловом уровне), так как потребуются очень большие словари исключений. В принципе можно задавать дополнительный признак вида, но это снижает автоматизированность модели и поэтому не проводилось в данной работе.

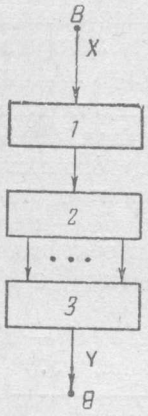

Рис. 1.

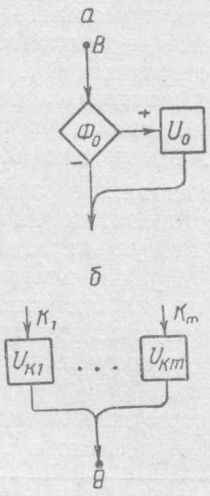

Рис. 2.

Составные блоки *1*, *2*, *3* (рис. 1) состоят из элементарных блоков — распознавателей и операторов. Распознаватель Φ_l проверяет, удовлетворяет ли слово, поданное на его вход, какому-либо условию. В случае выполнения этого условия слово выходит из распознавателя по стрелке, отмеченной знаком плюс, при невыполнении — по стрелке со знаком минус. Блок U_p, который каким-то образом изменяет поступающий на его вход сигнал, т. е. выполняет над этим сигналом определенную операцию, является оператором.

Опишем функционирование составных блоков алгоритмов на уровне их элементарных распознавателей и операторов. Блок нормализации *1* (рис. 2, *а*) не является обязательным и введен для упрощения решения. Он состоит из распознавателя Φ_0, проверяющего конец слова на *ся* или *сь*, и оператора U_0, отбрасывающего две последние буквы слова. Если входная глагольная форма имеет постфикс *ся* или *сь*, то блок *1* его отбрасывает. Блок классификации *2* (рис. 3) является основным блоком

алгоритма A_k. Он состоит только из распознавателей, поэтому не изменяет поступающий с выхода блока *1* сигнал, а лишь направляет его на один из m пронумерованных выходов, каждый из которых соединен с соответствующим (и таким же образом пронумерованным) входом блока *3*. Выходной блок *3* (рис. 2, *б*) состоит из m параллельно соединенных операторов замены слова на выходной сигнал-признак. Каждому сигналу y_{kj} (см. таблицу y_{kj}) соответствует оператор U_{kj}, который формирует этот выходной сигнал.

Рассмотрим подробнее блоки классификации, входящие в состав алгоритмов $A_1 — A_4$.

Блок *2* алгоритма A_1 (рис. 3, *а*) состоит из распознавателя Φ_{11}, который проверяет конец слова на одну из букв *е, м, т, ь, у* или *ю*. Блок классификации алгоритма A_2 (рис. 3, *б*) также можно получить из одного распознавателя Φ_{21}, проверяющего конец слова на *ем, им, ут, ют, те*. Распознаватели Φ_{22} и Φ_{23} вводятся для получения точного решения. Блок Φ_{22} проверяет слово на совпадение с глаголами *есть* и *надоесть* (и их производными),

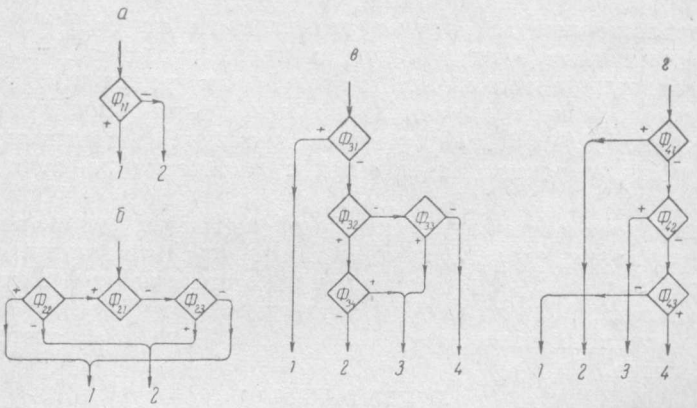

Рис. 3.

стоящими в первом лице единственного числа, а Φ_{23} — с глаголом *суть*. Введение в блок классификации алгоритма A_3 (рис. 3, *в*) распознавателя Φ_{34}, сравнивающего слово на совпадение с глагольной формой *есть* или *суть*, позволяет получить точное решение задачи 3. Распознаватели, обязательные для ее решения, проверяют последнюю букву слова на одну из букв *м, у, ю* (Φ_{31}); *е, ь* (Φ_{32}); *т* (Φ_{33}). На рис. 3, *г* показан блок *2* алгоритма A_4, в состав которого входят три элементарных распознавателя. Блоки Φ_{41} и Φ_{42} проверяют конец слова соответственно на *ла* и *ло*; Φ_{43} и распознаватель Φ_{11}, описанный выше, однотипны. Алгоритм A_4 аналогичен модели определения рода глаголов русского языка, описанной в работе [3], и немного проще этой модели.

Алгоритмы автоматического морфологического анализа категорий лица и числа глаголов суженной парадигмы значительно проще, чем алгоритмы анализа тех же категорий всего множества глагольных форм. В качестве примера можно сослаться на алгоритм морфологической классификации глаголов по признаку лица [4]. Если множество входных слов расширить глаголами в инфинитиве, то предлагаемые алгоритмы изменятся совсем незначительно. Алгоритм различения неопределенной и личной формы глагола аналогичен модели [5], но менее сложен.

Если необходимо одновременно решать все четыре задачи, желательно располагать обобщенным алгоритмом анализа, так как он получается более компактным, чем простое объединение *алгоритмов* $A_1 — A_4$. Обобщенный алгоритм определения грамматических категорий времени, числа, лица и рода глаголов суженной парадигмы можно представить в виде рис. 1. Математические модели нахождения значений лица и числа глаголов повелительного наклонения и некоторые другие также получены в виде, показанном на рис. 1. Это подчеркивает целесообразность подобного представления моделей.

Предложенные алгоритмы реализованы на ЭЦВМ. Проводятся машинные эксперименты в целях исследования уровня достоверности предсказания моделей для различных текстов. Эти исследования основаны на применении техники статистических вычислений [6]. Полученные значения вероятностей достоверного предсказания моделей практически не отличаются от единицы.

Описанные алгоритмические модели могут использоваться для дальнейших исследований, а также играть самостоятельную роль как математические описания некоторых психических функций человека. Модели могут найти применение в различных системах автоматического морфологического анализа текстов русского языка.

ЛИТЕРАТУРА

1. Грамматика современного русского литературного языка. М., «Наука», 1970. 767 с.
2. Орфографический словарь русского языка, изд. 11-е. М., «Сов. энциклопедия», 1971. 520 с.
3. Шабанов-Кушнаренко Ю. П., Бондаренко М. Ф., Соловьева Е. А. Математическая модель определения грамматической категории рода глаголов русского языка.— В сб.: Проблемы бионики. Вып. 11. Харьков, 1973, с. 3—5.
4. Шабанов-Кушнаренко Ю. П., Соловьева Е. А. Математическая модель принятия решений при классификации глаголов по признаку лица.— В. сб.: Проблемы бионики. Вып. 11. Харьков, 1973, с. 139—142.
5. Бондаренко М. Ф., Соловьева Е. А. Методы решения задач морфологической и субморфологической классификации.— В сб.: Проблемы бионики. Вып. 10. Харьков, 1972, с. 145—150.
6. Миропольский А. К. Техника статистических вычислений, изд. 2-е. М., «Наука», 1971. 576 с.