МЕТОДЫ РЕШЕНИЯ ЗАДАЧ МОРФОЛОГИЧЕСКОЙ И СУБМОРФОЛОГИЧЕСКОЙ КЛАССИФИКАЦИИ

М. Ф. Бондаренко, Е. А. Соловьева

Человек способен классифицировать части речи или части слов по некоторым признакам, т. е. решать задачи морфологической классификации. Нас интересуют математические модели такой способности абсолютно грамотного человека, выполняющего предложенную задачу классификации точно в соответствии с правилами грамматики русского языка [1, 2]. В данной работе рассмотрим задачу классификации глаголов на разряды по признаку формы (неопределенной или личной). На примере решения этой задачи продемонстрируем достоинства и недостатки предложенных ниже методов решения тех задач морфологической классификации, которые на основании одних только формальных признаков невозможно решить точно.

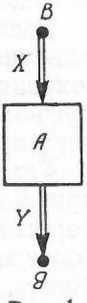

Рис. 1.

I. Решение задачи основывается на формальных признаках, за исключением случаев, когда это может привести к ошибкам. В таких случаях допускается минимально возможный словарь исключений, который дает возможность получить точное решение.

II. Решение задачи основывается только на формальных признаках, и потому оно получается с некоторой погрешностью.

Задачу можно также решать методом, представляющим собой комбинацию методов I и II.

При решении задачи классификации глаголов по признаку формы абсолютно грамотному человеку предъявляется глагол (например, из словаря русского языка [3]) в любой простой форме (одно слово). Испытуемый должен определить форму предложенного глагола. Необходимо математически описать такую психическую функцию человека.

Постановка задачи в общем виде представлена на рис. 1, где $X = \{x_1, \ldots, x_n\}$ — множество входных сигналов; $Y = \{y_1, \ldots, y_n\}$ — множество выходных сигналов; A — преобразователь информации, классифицирующий глаголы по признаку формы (неопределенной или личной). Подав на вход B преобразователя информации A какое-либо входное слово $x_i \in X$, на выходе g получим это же слово, классифицированное по признаку формы, т. е. выходной сигнал $y_i \in Y$, после чего на вход можно подавать новое слово. Необходимо составить математическое описание для A. В данной статье приведены два таких математических описания в виде алгоритмов AI (рис. 2) и AII (рис. 3).

Неопределенная форма глагола (инфинитив) рассматривается как начальная, исходная форма для всей системы глагола [1]. Поэтому представляется важным решение задачи выделения

глаголов в данной форме из всей массы глагольных форм. Результаты психологических экспериментов, а также анализ парадигмы глагола показали, что на основании одних только формальных признаков невозможно точно решить поставленную задачу. При этом было установлено, что все глагольные формы можно подразделить на три типа:

а) глагольные формы, классифицирующиеся на основании формальных признаков;

б) глагольные формы, классифицирующиеся при условии известности их лексических значений;

в) глагольные формы, классифицирующиеся только при наличии контекста.

Разделение на типы проведено для глагольных форм при решении конкретной задачи. Несмотря на это, такое разделение сохраняется в общем виде и для других задач классификации, причем множества слов одного или двух каких-либо типов в некоторых случаях могут оказаться пустыми.

Глагольные формы типа а) имеют различные формальные признаки для различных разрядов, на которые эти глагольные формы классифицируются. К типу а) в данной задаче принадлежит большая часть глагольных форм. К типу б) относятся те глагольные формы, которые, несмотря на одинаковые формальные признаки, принадлежат к различным разрядам. Например, все глаголы в неопределенной форме оканчиваются на «*ть*», «*ти*» или «*чь*». В то же время некоторые глаголы в личной форме (повелительное наклонение второе лицо единственное число) оканчиваются точно так же, причем иных формальных различий все вышеперечисленные глаголы не имеют. Ясно, что классифицировать такие глаголы на основании одних формальных признаков невозможно.

Глаголы типа в) при отсутствии контекста можно отнести к различным разрядам. Например, глагол *расти* относится одновременно и к неопределенной форме, и к личной. Так как слова предъявляются человеку без контекста, то в случае глагольных форм типа в) такой двоякий ответ для данной задачи будем считать верным.

Человек в состоянии точно решить задачу, если ему известны лексические значения глаголов типа б) и в), причем для поставленной задачи классификации число таких глаголов значительно. Следовательно, при решении данной задачи на ЭЦВМ необходимо «ознакомить» машину с большим числом глаголов, что нецелесообразно. Поэтому для решения применим методы I и II, исключающие такую необходимость.

Решение поставленной задачи методом I в виде блок-схемы алгоритма AI показано на рис. 2. В множество входных сигналов X включены все глаголы русского языка, входящие в словарь [3], и все простые формы этих глаголов. Множество выходных сигналов Y — это множество X, классифицированное по признаку формы, т. е. выходной сигнал y_i — это входной сигнал x_i с до-

бавлением признака формы: $П_1$, если глагол стоит в неопределенной форме; $П_2$ — если в личной; $П_1 \wedge П_2$, если без контекста глагол можно отнести и к неопределенной, и к личной формам.

Рассмотрим функционирование элементарных блоков алгоритма — распознавателей и операторов (общий принцип их действия описан в работе [4]). Распознаватели, входящие в состав алгоритма AI, выполняют проверку следующих условий: $Ф_1$ проверяет две последние буквы слова на «*ся*» или «*сь*»; $Ф_2$, $Ф_3$ и $Ф_4$ — те же буквы соответственно на «*чь*», «*ти*» и «*ть*»; $Ф_5$ — третью от конца букву слова на «*з*» или «*с*»; а $Ф_6$ — ту же букву на «*й*». $Ф_{И_k}$ ($k = \overline{1,5}$) проверяет, оканчивается ли входное слово на одно из слов словаря $И_k$, а $Фс_l$ ($l = \overline{1,6}$) — совпадает ли входное слово с одним из слов словаря $С_l$.

Операторы в алгоритме AI выполняют следующие действия; U_1 отбрасывает две последние буквы слова; $И_{П_1}$ заменяет слово, поданное на его вход, признаком $П_1$; $И_{П_2}$ — признаком $П_2$, а $И_{П_1 \wedge П_2}$ — признаком $П_1 \wedge П_2$. Каждая из цифр *1*, *2*, *3* на выходе любого блока означает, что этот выход должен быть соединен соответственно с оператором $И_{П_1}$, $И_{П_2}$, $И_{П_1 \wedge П_2}$.

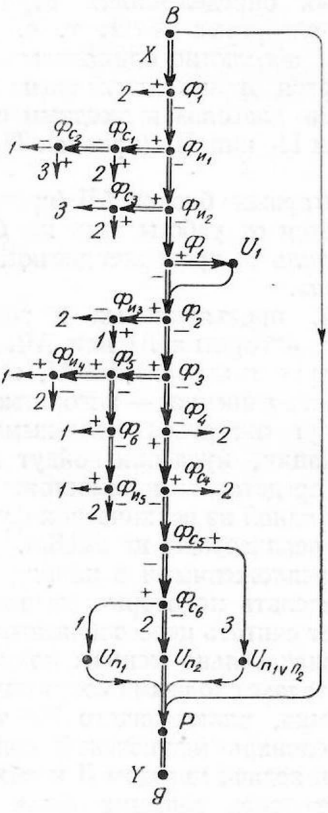

Рис. 2.

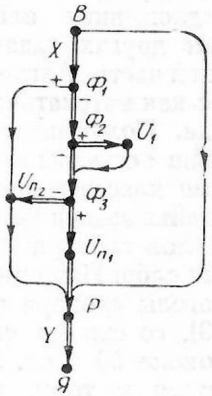

Рис. 3.

Блок P выполняет одновременно функции оператора и блока памяти. Запоминая поданное на его вход слово, блок P приписывает к этому слову поданный через некоторое время признак. Затем полученный сигнал поступает на выход алгоритма, после чего блок освобождается, и на вход алгоритма можно подавать новое слово.

На рис. 3 показана блоксхема алгоритма AII, который является решением задачи по методу II. Входные сигналы AII аналогичны входным сигналам AI, но их количество для AII может неограниченно возрастать за счет добавления новых слов, не входящих в состав словаря [3]. В зависимости от того, какие слова добавятся, может несколько меняться погрешность алгоритма A II, но на самом деле эти изменения не будут особенно заметными. В то же время нет практической необходимости расширения множества входных слов новыми глаголами, так как это множество составляют все глаголы современного русского языка со своими формами. Поэтому для определенности ограничим множество входных сигналов X алгоритма A II, т. е. будем считать входные сигналы AI и AII абсолютно одинаковыми.

Выходные сигналы AII отличаются от выходных сигналов AI только тем, что при классификации глаголов к входным словам могут добавляться только признаки Π_1 или Π_2 (признак $\Pi_1 \wedge \Pi_2$ исключается).

Функционирование всех элементарных блоков AII (кроме распознавателя Φ_3) ничем не отличается от работы этих же блоков в AI, описанных ранее; распознаватель Φ_3 проверяет две последние буквы слова на «*чь*», «*ть*» или «*ти*».

В зависимости от требований, предъявляемых к той или иной задаче, возможно применение алгоритма AI или AII. Когда необходимо точное решение, пользуются алгоритмом AI, если же требуется простота и большая скорость решения, — алгоритмом AII.

Предложенные алгоритмы могут оказаться полезными при решении других задач классификации, куда они войдут в виде составной части. Алгоритмы могут представить и самостоятельный интерес как математические модели одной из психических функций человека. Полученные алгоритмы реализованы на ЭЦВМ.

Решив поставленную задачу предложенными в начале статьи методами классификации, можно сделать некоторые выводы.

Решение задачи методом I следует считать целесообразным, если число слов словаря исключений значительно меньше количества входных слов. Например, в данной задаче входными могут являться все глаголы словаря русского языка, включающего 104 тысячи слов [3], со своими формами, а словарь исключений содержит всего около 50 слов. При решении задачи методом II необходимо стремиться к тому, чтобы погрешность решения была минимальной. В нашей задаче, решенной методом II, погрешность приблизительно равна 0,1%. В работе [5] математическая модель способности человека проверять правильность переноса слов, составленная методом II, также получена с небольшой погрешностью (менее 1%),

Выбор того или иного метода либо обоих методов одновременно зависит от конкретной задачи и результатов, которые небходимо получить. Метод II, например, обычно более прост, но всегда содержит некоторую погрешность. Когда необходимо абсолютно

точное решение, применяется метод I. Если же получение решения этим методом почему-то затруднительно или основным требованием является простота решения, то следует отдать предпочтение методу II. Метод II можно применять, если погрешность решения этим методом не превышает допустимую.

ЛИТЕРАТУРА

1. Грамматика русского языка, т. I. М., изд-во АН СССР, 1960.
2. Грамматика современного русского литературного языка. М., «Наука», 1970.
3. Орфографический словарь русского языка, изд. 11-е. М., «Сов. энциклопедия», 1971.
4. Л. А. Калужнин. Об алгоритмизации математических задач. Сб. «Проблемы кибернетики», вып. 2. М., Физматгиз, 1959.
5. Е. А. Соловьева. Математическое описание способности человека анализировать правильность переноса слов. Сб. «Проблемы бионики», вып. 8. Изд-во Харьковск. ун-та, 1972.