УДК 62.506.2

В. М. БОНДАРЕВ

АЛГЕБРАИЧЕСКОЕ ОПИСАНИЕ ЗАДАЧ МОРФОЛОГИЧЕСКОЙ ОБРАБОТКИ СЛОВ

Известно, что способность к переработке текстовой информации является важнейшим свойством человеческой психики. Для познавания и моделирования этого свойства представляется целесообразным изучить процессы грамматической обработки отдельных слов языка. Цель данной работы — попытка единым образом подойти к решению таких задач морфологии как словоизменение, грамматический анализ, поиск словарной формы слова, а также дать математическое описание процесса решения этих задач.

Введем понятие морфологической функции [1]. Назовем морфологической функцией $(n+2)$-местный предикат, принимающий значение 0 и 1. Обозначим его следующим образом: $L(x, y, z_1, z_2, \ldots, z_n)$. Аргументами этого предиката являются: x — некоторое слово; y — какая-либо форма слова; $z_1, \ldots, z_n$ — грамматические признаки формы слова. Значение функции L равно нулю, если заданы внутренне противоречивые (несовместные) значения ее аргументов. В противном случае значение функции L равно единице. Например, L («дорогой», «дорогим», дательный, женский, множественное) = 1, а L («дорогой», «дорогим», дательный, женский, единственное) = 0.

Введенный математический объект по праву называется функцией, так как осуществляет однозначное отображение области значений аргументов на множество $\{0, 1\}$. В самом деле, из языковой практики следует, что полный набор значений аргументов не может быть совместным и несовместным одновременно. Завершая описание морфологической функции, укажем, что под словом x и формой слова y мы будем понимать любую цепочку букв русского алфавита, число букв в которой не превышает некоторого наперед заданного числа A.

Для ответа на вопрос, каким образом в терминах морфологической функции будут выглядеть задачи морфологии, перечисленные выше, рассмотрим уравнение $L(x, y, z_1, z_2, \ldots, z_n) = 1$, которое назовем морфологическим уравнением. Ясно, что этому уравнению будут удовлетворять лишь такие $x, y, z_1, z_2, \ldots, z_n$, которые совместимы друг с другом. Зададим некоторые значения переменных $x, z_1, z_2, \ldots, z_n$ и решим морфологическое уравнение относительно неизвестной y. В результате мы совершим переход от слова и грамматических признаков к соответствующей форме слова. Зададим теперь значение y и решим морфологическое уравнение относительно неизвестных $z_1, z_2, \ldots, z_n$. При этом мы по известной форме слова определим соответствующие ей грамматические признаки. Наконец, зададим конкретную форму слова y и решим морфологическое уравнение относительно неизвестной x. В результате мы получим слово, соответствующее одной из его форм. Таким образом, с помощью морфологической функции легко формулируются три задачи, указанные нами ранее. Очевидно, что существует 2^{n+2} задач, допускающих такую же формулировку. Три рассмотренные задачи входят в их число.

Попытаемся представить морфологическую функцию в виде суперпозиции некоторых строго определенных и достаточно простых функций [2]. Предварительно изменим обозначение переменных функции L. Дадим им сквозную нумерацию и обозначим единым символом x. Это связано с тем, что дальнейшие рассуждения будут симметричны относительно всех аргументов функции L. Поскольку область определения каждой переменной конечна, пронумеруем все значения переменной из области определения, и в выражение $x_i = j$ будем вкладывать следующий

смысл: переменная x_i принимает j-е значение из своей области определения.

Для каждого аргумента функции L определим класс K_i функций, которые назовем характеристическими. Индекс i означает, что класс K_i соответствует i-му по счету аргументу. Обозначим эти функции через x_i^j:

$$x_i^j = \begin{cases} 1, & \text{если } x_i = j, \\ 0, & \text{если } x_i \neq j \end{cases} \quad (i = 1, 2, \ldots, n+2). \tag{1}$$

Все функции класса K_i определены лишь на значениях переменной x_i, а их число совпадает с количеством различных значений соответствующей переменной.

Рассмотрим произвольный набор значений аргументов, обращающий функцию L в единицу. Определим значения характеристических функций на этом наборе. Ясно, что вся совокупность их значений однозначно определяется исходным набором значений аргументов, причем для различных исходных наборов соответствующие совокупности значений характеристических функций также будут различны.

Таким образом, мы как бы перешли от многозначного кодирования значений аргументов морфологической функции к их двоичному кодированию. Проделав описанный переход для каждого «единичного» набора значений переменных морфологической функции, получим табличное задание некоторой функции L_2 двоичной логики. Условимся, что она принимает значение «1» на всех тех наборах аргументов, которые получены нами из «единичных» наборов функции L. На всех остальных двоичных наборах такой же длины определим ее значение равным «0». Известным методом перейдем от табличного задания L_2 к ее СДНФ:

$$L_2 = \bigvee_{t=1}^{T} \bigwedge_{i=1}^{n+2} \left[\bigwedge_{j=1}^{m_i} (x_i^j)^{\sigma} \right]. \tag{2}$$

Индекс σ означает, что двоичная переменная x_i^j может быть взята со знаком отрицания или без него; T — число дизъюнктивных членов в формуле, или, что то же самое, число «единичных» наборов значений аргументов морфологической функции; m_i — число всевозможных значений переменной x_i.

Рассмотрим отдельно выражение, стоящее в квадратных скобках. Нетрудно понять, что только одна двоичная переменная может входить в это выражение без показателя отрицания. Действительно, если соответствующий аргумент x_i функции L принимает свое q-е значение, то в соответствии с (1) $x_i^q = 1$, $x_i^j = 0$ $(j = 1, 2, \ldots, q-1, q+1, \ldots, m_i)$. Зафиксируем этот факт:

$$\bigwedge_{j=1}^{m_i} (x_i^j)^{\circ} = x_i^q \wedge (\neg x_i^1 \wedge \neg x_i^2 \wedge \ldots \wedge \neg x_i^{q-1} \wedge \neg x_i^{q+1} \wedge \ldots \wedge \neg \neg x_i^{m_i}). \tag{3}$$

Но с другой стороны справедливо тождество

$$x_i^q = \neg(x_i^1 \vee x_i^2 \vee \ldots \vee x_i^{q-1} \vee x_i^{q+1} \vee \ldots \vee x_i^{m_i}), \qquad (4)$$

которое непосредственно отражает то, что переменная x_i может принять q-е или одно из остальных значений, причем один из исходов исключает другой. Преобразуем (4) по правилу де Моргана:

$$x_i^q = \neg x_i^1 \wedge \neg x_i^2 \wedge \ldots \wedge \neg x_i^{q-1} \wedge \neg x_i^{q+1} \wedge \ldots \wedge \neg x_i^{m_i}. \qquad (5)$$

Подставим выражение (5) в (3):

$$\bigwedge_{j=1}^{m_i} (x_i^j)^{\sigma} = x_i^q \wedge x_i^q = x_i^q, \qquad (6)$$

а затем (6) в (2):

$$L_2 = \bigvee_{t=1}^{T} \bigwedge_{i=1}^{n+2} x_i^{q_i}. \qquad (7)$$

Индекс, появившийся у показателя q, означает, что каждый аргумент принимает свое значение независимо от остальных. Если вспомнить о том, что двоичные аргументы функции L_2 сами являются функциями от многозначных аргументов x_i ($i = 1, 2, \ldots, n+2$) морфологической функции, то выражение (7) можно считать записью функции L в виде суперпозиции дизъюнкции, конъюнкции и характеристических функций:

$$L = \bigvee_{t=1}^{T} \bigwedge_{i=1}^{n+2} x_i^{q_i}. \qquad (8)$$

Множество характеристических функций, дизъюнкцию и конъюнкцию естественно назвать базисом, в котором записана функция L. Этот базис полон в том смысле, что в нем может быть записана любая морфологическая функция. Справедливость этого утверждения вытекает из самого процесса построения выражения (8).

Ранее говорилось, что количество характеристических функций в отдельном классе равно числу элементов области определения соответствующей переменной. Поэтому число характеристических функций в классе таких аргументов, как форма слова или его словарная форма, должно быть очень велико. Если существует отображение, позволяющее взаимно-однозначно сопоставить всякому набору значений переменных морфологической функции некоторый набор значений других переменных, число которых может быть больше, но области определения невелики, то можно воспользоваться этим отображением и определить функцию L на новом наборе переменных.

Примером такого отображения может служить разбиение русских слов на буквы с указанием для каждой буквы ее позиции в слове или разбиение слов на морфы с указанием для каждого

морфа его значения. В первом случае все слова следует дополнить пробелами до длины, которую имеет самое большое слово, и считать пробел особой буквой. Во втором — следует считать отсутствие некоторого морфа в слове его особым, нулевым значением. В обоих случаях уменьшается суммарное число функций в базисе, а главное, построенная в этом базисе функция приобретает «объяснительную силу». Другими словами, в своей структуре она может отражать связь между значениями морфов и грамматических признаков или между другими категориями морфологии.

Присутствие в базисе операций дизъюнкции и конъюнкции позволяет широко пользоваться известным тождеством алгебры логики

$$x \wedge y \vee x \wedge z = x \wedge (y \wedge z) \qquad (9)$$

в целях уменьшения размеров выражения, описывающего морфологическую функцию.

Кроме того, это позволяет разложить выражение для морфологической функции на независимые части, применив некоторый аналог известной в алгебре логики теоремы о разложении. По-видимому, только расчленив выражение (8) на более или менее простые части, зависящие от меньшего числа переменных, чем все выражение в целом, можно подойти к практической реализации морфологической функции, скажем, в виде программы для ЭВМ.

СПИСОК ЛИТЕРАТУРЫ

1. Математическое моделирование процессов грамматической обработки словоформ русского языка.— Тезисы VIII Всесоюзного симпозиума по кибернетике. Тбилиси, 1976, с. 524—526. Авт.: Т. А. Недзельская, А. Ф. Осыка, А. И. Чугун, Ю. П. Шабанов-Кушнаренко.
2 Шабанов-Кушнаренко Ю. П. Применение метода нуль-органа в лингвистике. См. статью в настоящем сборнике.

Поступила 9 марта 1977 г.