УДК 62.506.2

Э. М. БУЗНИЦКАЯ, Н. К. СВИНАРЬ

ОБ ОДНОМ ПОДХОДЕ К ЭКСПЛИКАЦИИ ПОНЯТИЯ ЕСТЕСТВЕННОГО ЯЗЫКА

С появлением ЭВМ возникла проблема общения человека с машиной. Человеку удобнее выражать мысли в языковой форме и, следовательно, надо обучить машину непосредственно воспринимать и перерабатывать устную речь и тексты на естественных языках человека. Эта сложная проблема пока еще не решена. Чтобы добиться успеха, нужно понять в первую очередь закономерности языка и уметь описывать их на том уровне логической глубины и полноты, который может дать только математика.

Цель наших исследований в области теории естественного языка состоит в математическом описании лингвистической системы, носителем которой является человек. По-видимому, имеет смысл изучать принципы функционирования этой последовательной системы, выделяя более простые ее структуры, содержательно соответствующие разным уровням естественного языка. Исследовать язык означает исследовать соответствующие функции. Задача состоит в том, чтобы указать области их опре-

деления, построить их в явном виде и описать их свойства, согласованные с психофизическими реакциями человека.

Введем в рассмотрение некоторые понятия. Пусть имеется множество A. Составим из его элементов $a_1, a_2, \ldots, a_n$ упорядоченную n-ку $\langle a_1, a_2, \ldots, a_n \rangle$, в которой одни и те же элементы на различных местах могут встречаться многократно. Множество всевозможных упорядоченых n-ок назовем n-й степенью множества A и будем обозначать через A^n. Образуем объединение всех таких множеств: $S = A^1 U A^2 U \ldots U A^n$. Введенное множество состоит из всевозможных упорядоченных конечных последовательностей элементов множества A. Пусть A — морфологический алфавит русского языка; n-ю степень морфологического алфавита назовем множеством *морфологических выражений* длины n и обозначим через S. Примеры морфологических выражений:

$\langle ррдзот \rangle$, $\langle неская \rangle$, $\langle красно \rangle$.

Множество S будем называть *словарем*. Рассмотрим характеристическую функцию f, выделяющую подмножество таких морфологических выражений, которые носитель языка будет рассматривать как допустимые в этом языке, хотя, быть может, совершенно ему непонятные и незнакомые. Примеры:

$\langle глокая \rangle$, $\langle вет\,кали \rangle$, $\langle анокон \rangle$, $\langle стол \rangle$.

Говоря о существовании такой функции, мы имеем в виду существование ее только в том смысле, что имеется носитель этой функции — человек. Если предъявить испытуемому любое морфологическое выражение из словаря S, то он сумеет выработать вполне однозначный ответ. Этот ответ можно рассматривать как строго детерминированный сигнал из множества $\{0,1\}$. Таким образом, можно определить подмножество $S' \subset S$, элементы которого в дальнейшем будем называть *псевдословами*.

Введем понятие грамматической категории, задающее разбиение множества S' на два пересекающихся подмножества Γ_i и $\bar{\Gamma}_i = S' \setminus \Gamma_i$. Характеристическую функцию, конструктивно задающую грамматическую категорию γ_i на множество S', можно записать в виде

$$\gamma_i(x) = \begin{cases} 1, & \text{если элемент } x \text{ обладает категорией } \gamma_i; \\ 0 & \text{в противном случае.} \end{cases}$$

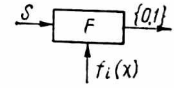
Последовательность из n грамматических категорий $\gamma_1, \gamma_2, \ldots, \gamma_n$ дает разбиение множества S' на 2^n *элементарных классов* (некоторые из них могут быть пустыми). Тогда любому элементу $x \in S'$ можно поставить в соответствие хотя бы один элементарный класс Ψ_k $(k = 1, 2, \ldots 2^n)$. Любое подмножество Ψ_k может быть получено как пересечение подмножеств, соответствующих грамматическим категориям и их отрицаниям:

$$\psi_k = \bigcap_{i \in M_k} \Gamma_i \bigcap_{j \in \bar{M}_k} \bar{\Gamma}_j \; (M_k \subseteq \{1, 2, \ldots, 2^n\}, \; \bar{M}_k = \{1, 2, \ldots, 2^n\} \setminus M_k).$$

Соответствующая характеристическая функция представима в виде

$$\psi_k(x) = \bigwedge_{i \in M_k} \gamma_i(x) \bigwedge_{i \in \overline{M}_k} \neg \gamma_j(x).$$

Две грамматические категории γ_{i_1} и γ_{i_2} назовем *совместимыми*, если $\exists x (\gamma_{i_1}(x) \text{and} \subset \gamma_{i_2}(x))$. Очевидно, что характеристическая функция любого непустого элементарного класса может быть представлена как конъюнкция характеристических функций совместимых грамматических категорий, которую в дальнейшем будем называть *логической формулой* элементарного класса.



Однако существуют такие элементы множества S', которым можно поставить в соответствие несколько элементарных классов, объединение которых будем называть *дистрибутивным классом*. Каждому элементу $x_l \in S'$ будет соответствовать единственный дистрибутивный класс Φ_l, логическая формула которого запишется в виде $\varphi_l(x) = V_k \psi_k(x)$. Все дистрибутивные классы содержатся во множестве всевозможных объединений элементарных классов. Введенное понятие дистрибутивного класса отражает явление омонимии, свойственное в той или иной мере естественным языкам.

Теперь перейдем к вопросам, связанным с изучением системы лингвистических функций человека при решении задач морфологической классификации. Общая схема исследования способности человека решать поставленные выше задачи по методу «черного ящика» следующая (рисунок).

Вход S соответствует множеству морфологических выражений, предъявляемых испытуемому. Вход $f_i(x)$ соответствует частной лингвистической функции, конкретный вид которой определяется логической формулой (i — номер или код задания, даваемого испытуемому). Лингвистическую модель, соответствующую языковому поведению человека при решении задач морфологической классификации, можно представить как универсальную для этой системы функций. Выход {0,1} соответствует результатам классификации всевозможных морфологических выражений в зависимости от $f_i(x)$.

Задача состоит в получении математического описания лингвистических функций, реализуемых человеком, в виде алгоритма, преобразующего входное множество S в выходное {0,1} в зависимости от $f_i(x)$.

Исходной информацией для построения такого алгоритма служит список неформальных грамматических правил, регулирующих ответ испытуемого. Пусть элементарный класс описывается логической формулой $\Psi_k(x)$. Необходимо найти все элементы этого класса из множества S'. Пусть Γ_{α_1}, Γ_{α_2},..., Γ_{α_k}

грамматические категории, входящие в данный элементарный класс. Тогда интересующее нас множество будет иметь вид

$$\psi_k = \bigcap_i \Gamma_{\alpha_i} (i = 1, 2, \ldots, k),$$

где $\Psi_k(x)$ — характеристическая функция множества Ψ_k, которую можно представить в виде

$$\psi_k(x) = \begin{cases} 1, & \text{если } x \in \psi_k; \\ 0, & \text{если } x \notin \psi_k. \end{cases}$$

Пусть $\gamma_i(x)$ $(i = \alpha_1, \alpha_2, \ldots, \alpha_k)$ — характеристические функции множеств Γ_i с подобными значениями. Тогда $\psi_k(x) = \bigwedge_i \gamma_{\alpha_i}$.

Пусть функция $\gamma_i(x)$ задана системой правил $P_{i1}, P_{i2}, \ldots, P_{im_i}$. Эти правила задают систему характеристических функций $\beta_{i1}, \beta_{i2}, \ldots, \beta_{im_i}$ на множестве S'. Эти функции определяют некоторые множества $B_{i1}, B_{i2}, \ldots, B_{im_i}$. Если система правил полна, то $\gamma_i(x) = \beta_{i1}(x)$ and $\beta_{i2}(x)$ and ... and $\beta_{im_i}(x)$. Система правил может быть неполной, тогда $\Gamma_i \neq B_{i1} \cap B_{i2} \cap \ldots \cap B_{im_i}$, но $\Gamma_i \subset B_{i1} \cap B_{i2} \cap \ldots \cap B_{im_i}$. Формальным эквивалентом системы правил, записанных на естественном языке, может быть язык математической логики. Тогда система правил есть конъюнкция некоторого числа высказываний.

Последний этап решения задачи состоит в создании программы, реализующей алгоритм данной лингвистической функции на ЭВМ.

СПИСОК ЛИТЕРАТУРЫ

1. Маркус С. Теоретико-множественные модели языка. М., «Наука», 1970. 332 с.
2. Клини С. Математическая логика. М., «Мир», 1973. 480 с.

Поступила 20 июля 1975 г.