УДК 510.62

З. В. ДУДАРЬ, В. С. МАКСИМОВА

МОДИФИЦИРОВАННЫЙ МЕТОД ВЕТВЕЙ И ГРАНИЦ ДЛЯ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ЛИНГВИСТИЧЕСКИХ ЕДИНИЦ

Язык — это перархическая многоуровневая система лингвистических единиц (букв, леорфов, слов, их сочетаний, предложений и т. д.). Единицы данной системы языка (русского) вступают во взаимодействие (ансамбль) с другими системами (языками) и уровнями систем. Результат этого взаимодействия — проникновение иноязычных слов, появление новых закономерностей в лингвистических описаниях. ,

Лингвистические единицы, вступающие в ансамбль с единицами различных уровней языка, отвечающие законам синтагматики, динамики развития языка, обладающие семантической перспективой, называются потенциально-возможными сочетаниями. Потенциально-возможные сочетания с максимальной частотностью явления в языке образуют ядро языка.

Лингвистические единицы, не вступающие в ансамбль с другими единицами, не отвечающие законам синтагматики и динамики развития языка, не обладающие семантической перспективой, имеющие нулевую частотность в языке, называются потенциально-невозможными сочетаниями. Эта область языка является тупиковой, т. е. не имеющей перспективы сочетаний и развития ядра.

Тупиковые области и области с минимальной частотностью явлений в языке образуют периферийные области в языке. Известно, что только 30 % от потенциально-возможных сочетаний лингвистических единиц различных уровней вступают в ансамбль друг с другом, встречаются в естественном языке, поэтому задача выделения ядра языка, прогнозируемых периферийных и тупиковых областей имеет особый практический смысл для многих задач искусственного интеллекта, автоматической обработки информации и речи.

Наиболее эффективный метод общения с высокопроизводительными ЭВМ —непосредственное обращение к внутренним машинным операциям с помощью естественной человеческой речи. В этой связи одной из актуальнейших задач является моделирование лингвистических закономерностей процессов, распознания и синтеза речи. Закономерности синтагматики русской речи являются основополагающими в этих исследованиях: они позволяют прогнозировать следование сочетаний лингвистических единиц в языке. Таким образом, используя результаты исследований синтагматики языка, мы будем двигаться не слепым перебором по комбинаторным ансамблям дискретных лингвистических единиц (КАДЛЕ), а руководствоваться мощным критерием следования — допущениями и ограничениями языка. Естественно, что практические ограничения синтагматики, т. е. запреты на сочетаемость и допущения сочетаемости в естественном языке, существенно сокра-

ищают область перебора, объем памяти и время пользования ЭВМ, однако увеличивают вероятность предсказания явления. Существенно ограничат перебор КАДЛЕ и наиболее «информативные», т. е. множество потенциально-возможных сочетаний и их подмножеств, участвующих в процессах естественного языка с максимальной частотностью явления.

Области запретов и допущений сочетаний разделяют введением системы критериев. Предлагается система критериев движения по графу КАДЛЕ. КАДЛЕ — система многокомпонентная, обладающая системой «вращения» в структуре КАДЛЕ и обеспечивающая многоуровневую, многомерную и многокомпонентную связь в КАДЛЕ. Деревья КАДЛЕ построены на основании процесса структурирования (расширения) ядер в КАДЛЕ, а прохождение по ним к цели — на основании критериев следования: критерия движения по вершинам КАДЛЕ в соответствии с некоторой оценочной функцией $K_д$ критерия комбинаторной мощности КАДЛЕ определенного уровня $K_м$; критерия позиционной активности КАДЛЕ $K_А$; критерия структурного взаимодействия КАДЛЕ различных уровней $K_{вз}$.

Критерий оценки должен учитывать перспективность появления новых вершин дерева КАДЛЕ совместно с величиной частотного явления. Эвристические методы достижения цели позволяют КАДЛЕ представить в виде дерева целей.

Критерием оценки наиболее «информативных», т. е. потенциально возможных сочетаний максимальной частотности, является функция f, значение которой $f(n)$ для любой вершины n, представляет собой сумму оценок КАДЛЕ от ядра S (начальной вершины) сочетаний до целевой вершины n, т. е. конечного, обладающего семантикой множества сочетаний.

Таким образом, $f(n)$ — критерий оценки перебора по вершинам G при условии, что этот перебор начинается с ядра S и проходит через вершину n_i. По этому предположению движение должно осуществляться в сторону максимальной величины вероятностных частот сочетаний КАДЛЕ.

Иллюстрируя изложенное выше предположение, введем некоторые обозначения. Пусть функция критерия оценки $R(n_i, n_j)$ дает действительную оценку пути КАДЛЕ между двумя вершинами n_i и n_j по максимальной величине вероятностных частот КАДЛЕ. Функция критерия оценки не определена для вершин (сочетаний КАДЛЕ), между которыми нет пути, т. е. множеств КАДЛЕ, которые образуют область тупиковых вершин — запретов естественного языка, частотность которых равна 0.

Если T — множество целевых вершин, т. е. потенциально-возможных сочетаний КАДЛЕ, обладающих потенцией производить удлинение множеств сочетаний и семантической перспективой, то путь перебора от вершины до цели обозначим через

$$h(n_l) = \max_{n_j \in T} R(n_l, n_j).$$

Путь от вершины ядра КАДЛЕ n_i к целевой вершине n_j, для которого достигается $h(n_i)$, — оптимальный.

Оценка оптимального пути $k(S, n)$ от вершины ядра S КАДЛЕ до некоторой произвольной целевой вершины n, т. е. до потенциально-возможного сочетания КАДЛЕ максимальной частоты и семантической потенции, определяется при $q(n)=R(S, n)$ для всех n, достижимых из ядра S.

Тогда функция критерия оценки наиболее информативных по величине максимальной частотности явления КАДЛЕ — $f(n)$ для любой вершины n равна сумме действительных оценок веса оптимального пути перебора синтагматических закономерностей КАДЛЕ от ядра S до вершины n и веса оптимального пути от вершины n до любой из целевых вершин $f(n)=q(n)+h(n)$.

Следовательно, значение $f(n)$ — вес оптимального пути перебора КАДЛЕ при условии, что он проходит через вершину n. В случае, когда $f(S)=h(S)$, действительный вес пути перебора КАДЛЕ движется от ядра S к целевой вершине, т. е. к конечному множеству потенциально-возможных сочетаний КАДЛЕ без ограничений.

Критерий оценки движения по дереву перебора КАДЛЕ введем с помощью оценочной функции $\hat{f}(n)=\hat{q}(n)+\hat{h}(n)$, где $\hat{q}$ — оценка для q, $\hat{h}$ — оценка для h.

В качестве $\hat{q}(n)$ выбираем вес пути прохождения перебора КАДЛЕ от ядра S до n целевой вершины. Это достигается суммированием весов максимальных частот сочетаний КАДЛЕ, лежащих на пути прохождения. Таким образом, передача управления движения синтеза сочетаний КАДЛЕ идет в сторону максимальной величины $\hat{q}(n_i)$.

Применение формального аппарата универсальной алгебры конечных предикатов [1] — алгебры отношений единиц естественного языка — позволит формализовать явления максимальной частотности КАДЛЕ, детально и конкретно описать явления более интенсивные по частотности

Метод ветвей и границ [2] хорошо разработан, является классическим для задач искусственного интеллекта. Его новая модификация [3] применена к задачам анализа синтеза и распознавания речи. При решении этих задач учитываются законы синтагматики, парадигматики и динамики комбинаторных ансамблей дискретных лингвистических единиц естественного языка, т. е. потенциально возможных и потенциально невозможных сочетаний графем для письменной речи, сочетаний звуков на фонемном уровне, обладающих потенцией производить или не производить удлинение множеств сочетаний и возможной семантической перспективой.

Начальное состояние исследуемого подмножества сочетаний КАДЛЕ, т. е. так называемую «точку отсчета» рассматриваемого

подмножества сочетаний лингвистических единиц (ЛЕ), примем за начальную вершину графа G. Граф G представляет собой множество вершин, описывающих пространство состояний множества сочетаний КАДЛЕ языка и речи. Вершины — потенциально возможные и потенциально невозможные K-символьные сочетания КАДЛЕ, которые могут обладать потенцией удлинения и семантической перспективой. Пары вершин, связанные между собой законами динамики развития естественного языка, соединяются дугами, направленными от одного члена пары к другому. Таким образом, имеющийся граф G — направленный граф. Если дуга устремилась от вершины n_i к вершине n_j, то вершина n_j называется дочерней вершиной для n_i, а вершина n_i для n_j — родительской. Ситуация, когда вершины n_i и n_j будут «взаимнодочерними», описывает множества n-потенциально возможных сочетаний, обладающих потенцией удлинения сочетаний в обе стороны. В этой ситуации пара направленных дуг называется ребром графа G. Дуги графа представляют собой специальные операторы, преобразующие пространство состояний множеств сочетаний КАДЛЕ — вершин n_i в другое состояние, т. е. в n_j. Множества специальных операторов, обладающих определенными управляющими признаками, которые отвечают поставленной цели, называются специальными операторами цели. Обозначим их через Г. Далее, при воздействии на пространство структурированного ядра КАДЛЕ, находящегося в начальном состоянии, специальным оператором цели Γ_i строим следующие дочерние вершины графа G.

Предположим, начальная вершина G — 2-символьное ядро {ск} 4-х символьных сегментов {скаъ, скаа, ская, ...}. Структурирование ядра — приращение количества возможных подмножеств КАДЛЕ, обладающих аналогами в естественном языке и отвечающих законам синтагматики и динамики развития КАДЛЕ в языке. Множества структур сегментов удлиняются, если они имеют максимальную частотность явления. Наиболее интенсивный по частотности явления пучок сегментов образует новое ядро {ски}. Структурированное ядро более мощное по количеству входящих в него подмножеств символов языка, семантической перспективе, интенсивнее по частотности явления.

Подмножества {{ски} $\vee$ х е й}} представляют собой наиболее интенсивный пучок отношений G. Разработанный критерий оценки явления $\hat{q}(n)$, учитывающий перспективность появления новых вершин дерева КАДЛЕ на основании вероятностных и частотных явлений, условно разделит рассматриваемое множество структур ядра {ск} на подмножества потенциально возможных и потенциально невозможных:

1 подмножество {{ск} $\vee$ {а} $\vee$ {ъ $\vee$ а $\vee$ я}};

2 „ {{ск} $\vee$ {и} $\vee$ {х $\vee$ е $\vee$ й}};

3 „ {{ск} $\vee$ {о} $\vee$ {й}}.

Величина функции критерия оценки данных подмножеств равна:

$$1 \text{ подмножества } \hat{q}_1(n_i) = 78;$$

$$2 \quad \text{„} \quad \hat{q}_2(n_i) = 405;$$

$$3 \quad \text{„} \quad \hat{q}_3(n_i) = 150.$$

Как видно, именно формализация второго подмножества, частота встречаемости которого максимальна, была бы наиболее правомерна. Формализацию остальных пучков сегментов ядра следует воспринимать как исключение из правил, так как частотность язления минимальна.

Каждый из специальных операторов цели Γ_i обладает потенцией раскрытия S_i вершины, т. е. потенцией изменения состояния исследуемого подмножества сочетаний КАДЛЕ в другое, соответствующее цели. Следовательно, специальный оператор цели — функция, определенная на множестве состояний КАДЛЕ естественного языка или речи, принимающая значения из множества потенциально возможных сочетаний графем и фонем. Множества специальных операторов Γ_i, неопределенные на множестве состояний лингвистических единиц естественного языка и речи, образуют тупиковые области в структурах КАДЛЕ. Тупиковые области представляют собой множества потенциально невозможных сочетаний КАДЛЕ письменной и звучащей речи, не обладающих потенцией удлинения сочетаний КАДЛЕ, семантической перспективной и аналогами в естественном языке. Тупиковые области графа G являются областью тупиковых вершин — запретов КАДЛЕ звучащей и письменной речи, область частотности явлений которых минимальна или равна нулю.

Упорядоченная последовательность состояний множеств сочетаний КАДЛЕ — вершин графа n_{i_1}, n_{i_2}, ..., n_{iR}, в которой каждая вершина n_{ij} дочерняя для n_i, n_{j-1} (при $j=2, ..., R$) обладает потенцией удлинения множества, семантической перспективой, областью максимальной частотности явления, представляет собой структурированное ядро КАДЛЕ. Упорядоченная последовательность состояний множества сочетаний КАДЛЕ — вершин графа G, т. е. структурированное ядро, является также путем длины R от вершины n_{i_1} к вершине n_{iR}. Если существует путь от вершины графа G предыдущего символа сочетаний КАДЛЕ n_i к вершине n_j последующего символа сочетаний КАДЛЕ, то, следовательно, вершина n_j достижима из вершины n_i. От каждой предыдущей вершины графа G к последующей идут указатели, позволяющие от цели — конечного символа сочетаний КАДЛЕ, вернуться к начальной вершине и проследить путь решения. Каждая последующая вершина, описывающая пространство состояний множеств сочетаний КАДЛЕ, проверяется на соответствие законам естественного языка, динамике развития процессов синтеза и распознания речи.

Специальные операторы цели Γ_i, воздействующие на пространство состояний КАДЛЕ естественного языка, образуют решающую последовательность на пути прохождения КАДЛЕ по вершинам графа G, отвечающих поставленной цели. Множество вершин, соответствующих требованиям целевого состояния КАДЛЕ, образуют целевое множество вершин графа G. Целевое множество вершин графа G, описывающих пространство состояний КАДЛЕ, обладающих областью максимальной частотности явлений в естественном языке, образуют пространство КАДЛЕ письменной и звучащей речи. При наложении управляющих воздействий специальных операторов цели Γ_i на пространство целевого множества вершин графа G получим потенциально возможные сочетания КАДЛЕ, обладающие потенцией удлинения. Если полученные множества сочетаний КАДЛЕ отвечают законам естественного языка и речи, то можно предположить, что данными КАДЛЕ служат потенциально возможные слова или потенциально возможные сегменты слов.

Письменное или устное сообщение S, полученное при анализе, синтезе и распознании его элементов, представляет собой множество подмножеств сочетаний КАДЛЕ в виде

$$S = \sum_{i=n}^{n} S_i,$$

где S_i — вершина графа G. Последовательное применение специальных операторов цели Γ_i к $\{S_i\}$ вершинам, описывающим состояния подмножеств исследуемого множества КАДЛЕ, к последующим дочерним вершинам и т. д., дает представление графа G, определенное через Γ_i и $\{S_i\}$.

Рассмотрим пример, иллюстрирующий описания положений теоретико-графового метода описания КАДЛЕ.

Дано сочетание КАДЛЕ в виде письменного сообщения S, представленное как множество подмножеств: $S=\{\{э\}\ \{т\}\ \{а\}\ \{м\}\ \{а\}\ \{ш\}\ \{и\}\ \{н\}\ \{а\}\}$. Это же сообщение в задачах анализа, синтеза и распознавания звучащей речи обрастает многозначительностью, которая множится с невероятной скоростью, следовательно граф в случае, когда рассматриваются две ветви, приблизительно имеет вид:

т а м а ш и а

→ э н ,

д о б ы ж ы о

где →0 — начальная вершина графа G. Критерий оценки движения по дереву графа введен с помощью оценочной функции. В качестве $f(n)$ выбираем вес пути прохождения перебора КАДЛЕ от ядра S до S_i целевой вершины. Это достигается путем суммирования весов максимальных частот сочетаний КАДЛЕ, лежащих на пути прохождения. Передача управления движением синтеза и распознавания сочетаний КАДЛЕ идет в сторону максимальной вели-

чины S_i. Вершины {а} и {м} и т. д. являются дочерними для вершины {м}. Воздействуя на вершину {м} специальными операторами $\Gamma_а$, $\Gamma_ы$, $\Gamma_ъ$, получим потенциально возможные сочетания {м$\vee$а}=1, {м$\vee$ы}=1 и потенциально невозможные — {м$\vee$ы}=0, не отвечающие законам синтагматики естественного языка (русского). Следовательно, вершина {ы} в данном случае представляет собой тупиковую вершину из области запретов естественного языка. Частотность сочетаний {ш$\vee$ы} в письменной и звучащей речи равна нулю. Эта вершина не обладает потенцией удлинений множеств КАДЛЕ и семантической перспективой. В задачах синтеза и распознавания речи она участвовать не будет. Выделение области запретов естественной речи дает экономию при алгоритмизации процессов естественного языка и их обработки на ЭВМ. Множества вершин {а}, {ж}, {а}, {ш} являются «взаимно дочерними», так как обладают потенцией удлинения КАДЛЕ в обе стороны, отвечают законам естественного языка. Направление решения по вершинам графа G задано указателями. Рассматривая сочетания {ш$\vee$и}, {ш$\vee$ы}, фиксируем тупиковую вершину {ы}, не отвечающую законам синтагматики.

Приведенный пример слишком упрощенно рассматривает схему движения по вершинам графа G, описывающим пространство сочетаний КАДЛЕ. Это лишь попытка проиллюстрировать применение модифицированного метода ветвей и границ для изучения процессов перебора, поиска, анализа, синтеза и распознавания КАДЛЕ естественной речи.

Список литературы: 1. *Шабанов-Кушнаренко Ю. П.* Теория интеллекта. Математические средства. Х., 1984. 144 с. 2. *Алгоритм* решения задачи о коммивояжере/Дж. Литл, К. Мурти, Д. Суини, К. Кэрел//Экономика и мат. методы. 1965. № 1. С. 93—107. 3. *Максимова В. С.* Теоретико-графовый метод описания структурированных ядер естественного языка в интерактивных системах//Материалы пятой шк.-семинара «Интерактивные системы». Кутаиси, 2—10 апр. 1983 г. Тбилиси. 1983. С. 315—317.

Поступила в редколлегию 11.01.89