УДК 510.62

А. Ф. ОСЫКА, канд. техн. наук, *Н. И. БАРАНОВА*

О РАСПОЗНАВАНИИ ЭЛЕМЕНТОВ ЗНАЧЕНИЯ ТЕКСТА

Сообщение 1

Разработке систем обмена информацией с ЭВМ на языке человека уделяется большое внимание [1—3]. Необходимым условием построения таких систем является наличие простых и достаточно надежных процедур перехода от представления информации в виде предложений на входе ЭВМ к представлению на внутреннем языке системы.

Для машинных систем, обеспечивающих доступ к информации в ограниченной области знаний, предложен широкий спектр методов анализа: от представления запросов в виде наборов ключевых слов до получения графов синтаксических и семантических структур входных текстов [2, 4, 5]. В качестве примера рассмотрим два возможных запроса, поступающих на вход системы, которая выполняет функции кассира железнодорожной кассы.

«Я хочу купить купейный билет на поезд Харьков — Москва от Белгорода до Курска на сегодня».

«Можно ли купить купейный билет на поезд Харьков — Москва от Белгорода до Курска на сегодня?»

Наборы ключевых слов в обоих запросах одинаковы. Но реакция системы на эти запросы должна быть различной. Во втором случае достаточен ответ «да» или «нет», который можно получить при просмотре цифр остатков свободных мест на данный поезд. В первом случае требуется проверка наличия свободных мест и в положительном случае — поиск в базе данных и выдача сведений о конкретном месте. Следовательно, метод ключевых слов для анализа подобных запросов был бы недостаточен.

Вместе с тем явно нецелесообразно выполнять исчерпывающий анализ всего предложения или нескольких предложений запроса в синтаксических или семантических терминах. При анализе необходимо получить ответы из текста на некоторый список вполне конкретных вопросов типа: «О чем идет речь?» (покупка билета, получение информации о наличии мест, маршрут поезда, время прибытия поезда на станцию и т. п.), «Номер поезда?», «Станция отправления?» и т. д. Чтобы ускорить действие системы, ответы на подобные вопросы желательно получить без выявления структуры всего предложения запроса или ряда предложений.

Ответы содержатся в ключевых словах, но их семантическая роль может быть неоднозначной, так как отдельно взятое ключевое слово из запроса может быть правдоподобным отве-

том одновременно на несколько вопросов. В приведенных выше примерах запросов слово «Белгород», рассматриваемое изолированно, может быть ответом на вопросы: «Пункт отправления?», «Пункт прибытия?», «Начальный пункт маршрута поезда?», «Конечный пункт маршрута поезда?», «Промежуточный пункт маршрута поезда?». Это же относится и к словам «Курск», «Харьков», «Москва». Но выявление контекста (2—5 слов), характерного для той или иной семантической роли ключевого слова, зачастую полностью снимает подобную неоднозначность.

Рассмотрим один из способов учета диагностического контекста при выявлении семантических ролей ключевых слов в тексте запросов на естественном языке, вводимых в ЭВМ.

На этапе морфологического анализа каждой словоформе запроса (в том числе знакам препинания, дефисам и т. п.) приписывается следующая информация:

— номер словоформы в тексте;
— код семантического класса;
— часть речи;
— дополнительная морфологическая информация (род, число, падеж — для именных частей речи, инфинитив или время, число, лицо — для глаголов и т. п.);
— один или несколько номеров сетей переходов, предназначенных для выявления семантических ролей словоформ путем учета контекста. (Номера сетей приписываются не каждой словоформе, а только тем, которые хорошо предсказывают появление других слов в тексте).

Если словоформа лексически неоднозначна (многозначна, омонимична), то для каждого лексического значения выписывается полная информация как об отдельной словоформе со своими номерами сетей переходов. Но номер словоформы в предложении в этом случае повторяется. Морфологическая неоднозначность описывается в пределах единой информации о данном лексическом значении. Результаты морфологического анализа представлены в таблице.

№ словоформы	Словоформа	Семантический класс	Часть речи	Морфологические признаки	№ сети
1	Я	СУБ	Местоимение	1 л., ед. ч.	—
2	хочу	МОД	Глагол	1 л., ед. ч., н. в.	—
3	купить	ИМЕТЬ	Глагол	Инф.	—
4	купейный	ТМКУП	Прилагательное	М. р., ед. ч. и/в. п.	—
5	билет	МЕСТО	Существительное	М. р., ед. ч. и/в. п.	1, 2, 3, 4
. . .	. . . .	. . . .	. . . . .	. . . . .	. . .
16	сегодня	ДАТАI	Наречие	—	—
17	. . . .	ТЧК	Знак препинания	—	—

На втором этапе обработки в ЭВМ производится анализ контекста словоформ, выражающих существенные элементы значения запроса. Анализ выполняется с помощью сетей переходов, номера которых приписываются определенным словоформам при морфологическом анализе. В качестве примера

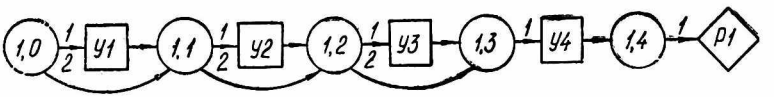

Рис. 1. Базовые типы структур

рассмотрим сеть № 1 (рис. 1), предназначенную для распознавания контекста словоформы «билет», который указывает на то, что речь идет о ситуации 1 — покупка билета (С1).

Любая сеть имеет узлы трех типов: круглые, прямоугольные и ромбовидные. Круглый узел 1.0 соответствует началу процедуры для выявления в тексте запроса словосочетания со значением С1. Остальные круглые узлы соответствуют словоформам запроса, которые выявляются данной сетью, т. е. удовлетворяют ее требованиям. Узлы этого типа не препятствуют продолжению анализа по любой из дуг, исходящих из данного узла. Сеть (рис. 1) предназначена для выявления словосочетаний, включающих в себя до четырех словоформ.

Квадратные узлы У1, У2, У3, У4 обозначают условия, которым должна удовлетворять словоформа запроса, чтобы она была выявлена сетью при переходе от одного круглого узла к другому. Каждое условие включает в себя все или некоторые из следующих характеристик: семантика словоформы, часть речи, необходимые морфологические признаки, место, где искать словоформу во фразе. В определенных случаях вместо семантики и морфологических признаков может указываться сама словоформа (например, предлог). Условие У1 в сети (рис. 1) имеет следующий вид: {Сем = СУБ, Мвп=1+5}. Это означает, что первая словоформа, выявляемая сетью, должна иметь семантический признак СУБ, а ее возможное место в предложении (Мвп) — от начала предложения (1) первые пять позиций. Условие У2 = {Сем = МОД, Мвп = 012}, т. е. семантический признак словоформы — МОД. Искать ее следует относительно ранее обнаруженной словоформы либо, если ее не было, относительно начала предложения (0), справа и слева (1) в глубину на две позиции (2). Условие У3 = {Сем = = ИМЕТЬ, Чр = глаг., Мвп = 012}. Смысл обозначений прежний. Добавляется признак части речи (Чр). Условие У4 = {Сем= =МЕСТО, Мвп=015}.

Дуги, соединяющие круглые вершины непосредственно, обозначают безусловный переход к следующему этапу анализа без выявления словоформы на данном этапе. Узел в виде ромба содержит результат, получаемый на выходе сети: вероятное

значение некоторого элемента смысла, существенного для понимания текстов в данной области знаний. На выходе одной сети могут получать значения несколько элементов смысла, каждый по одному значению. Результат $P1=\{SIT=C1\}$, т. е. элемент смысла «ситуация» (*SIT*) принимает значение «покупка билетов» (С1).

Процедура анализа, задаваемая сетью (рис. 1), проходит следующим образом. Порядок действий определяется порядком следования узлов в некотором пути сети, а также нумерацией дуг, исходящих из каждого узла. Цель анализа — найти во фразе словоформы, удовлетворяющие пути в сети, ведущему от вершины 1.0 к вершине результата. Процедура анализа начинается с узла 1.0, обращение к которому происходит после окончания морфологического анализа при считывании номера сети 1 в строке 5 (см. таблицу). При прохождении дуги 1.0.1 (дуги 1, исходящей из узла 1.0) проверяется условие У1, т. е. среди первых пяти словоформ запроса отыскивается словоформа с семантическим признаком СУБ. В данном запросе это словоформа «я». Во вспомогательный массив памяти производится запись «1, 1, Я, СУБ». Структура этой записи: «*j*, Н(*j*), Слов (*j*), Сем (*j*)», где *j* — номер словоформы в том порядке, в каком она была выявлена данной сетью; Н(*j*) — номер выявленной словоформы (см. таблицу); Слов(*j*) — словоформа, выявленная данным условием в сети; Сем(*j*) — семантический признак словоформы, который удовлетворил условию. При положительном результате проверки условия У1 осуществляется переход к узлу 1.1. В случае отсутствия словоформы, соответствующей условию У1, выполняется безусловный переход к узлу 1.1 по дуге 1.0.2. Производится запись «1, —, —, —».

Из узла 1.1 анализ продолжается по дуге 1.1.1. Проверяется условие У2, т. е. в таблице среди словоформ с номерами от 1—2 до 1+2 выявляется словоформа с семантикой МОД. (Отрицательные номера словоформ игнорируются). В данном запросе это словоформа «хочу». Выполняется запись «2, 2, хочу, МОД» и переход в узел 1.2. При невыполнении У2 производится безусловный переход по дуге 1.1.2 в узел 1.2 (например, в случае запроса: «Дайте мне билет...»).

Из узла 1.2 совершается переход по дуге 1.2.1 к проверке условия У3, для чего в таблице среди словоформ с номерами от 2—2 до 2+2 проверяется наличие глагола с семантикой ИМЕТЬ (в рассматриваемом запросе это «купить»). При выполнении условия У3 делается запись «3, 3, купить, ИМЕТЬ» и переход в узел 1.3. Если же такая словоформа не будет выявлена (условие У3 не выполнено), то по дуге 1.2.2 осуществляется безусловный переход в 1.3.

Узел 1.3 выполняет передачу управления на проверку условия У4, при этом в таблице среди словоформ с номерами от 3—5 до 3+5 выявляется словоформа с признаком МЕСТО

(в приведенном запросе это «билет»). В случае успешной проверки условия У4 делается запись «4,5, билет, МЕСТО» и переход в узел 1.4. Если условие У4 не выполняется, то осуществление данной процедуры прекращается без результата. Совершается переход к началу сети № 2, номер которой указан для словоформы «билет» (см. таблицу).

Из узла 1.4 выполняется переход к блоку формирования результата. В данном случае Р1 означает выдачу готовой информации: $SIT = C1$. Но в большинстве случаев в этом блоке заранее задана только переменная (левая часть равенства), а ее значение выбирается из соответствующих записей, сформированных при анализе сетью.

Нетрудно видеть, что сеть, представленная на рис. 1, дает возможность правильно анализировать запросы, в которых мысль о покупке билета выражена следующим образом: «Я хочу купить (1, 2, ...) билет(а)...», «Мне нужно приобрести (1, 2, ...) билет(а)...», «Хочу купить (1, 2, ...) билет(а)...», «Нуж(е)н(о) (1, 2, ...) билет(а)...», «Дайте мне (1, 2, ...) мест(о/а)...», «Мне нужно уехать (сегодня поездом № *X*)», «Дайте (1, 2, ...) плацкарт(а)...», «Мне (1, 2, ...) билет(а)...», «(Пожалуйста) (1, 2, ...) билет(а) (на поезд № *X* на сегодня)» и т. п. При этом допустимы практически любые перестановки словоформ в подобных фрагментах в начальной части фразы. Возможно наличие некоторых разделяющих слов.

Подобные фрагменты запросов будут правильно проанализированы в том случае, если на этапе морфологического анализа словоформам «я», «мне» и другим будет приписан семантический признак СУБ, словоформам «хочу», «нужно», «требуется», «необходимо» и т. п.— признак МОД, словоформам «купить», «приобрести», «получить» и т. п.— признак ИМЕТЬ, а словоформам «билет», «место», «плацкарт», «уехать» и т. п.— признак МЕСТО и отсылка к сети № 1.

Сеть № 2 предназначена для анализа словосочетаний в составе запроса со значением «тип места» (*TOP*). Она изображена на рис. 2. В этой сети У5={Сем=ТМ*ааа*, Мвп=1+15}, У6={Сем=МЕСТО, Мвп=012}, Р2={*TOP*=<Сем (1)>}.

Сеть № 2 имеет две особенности в отличие от сети № 1. В условии У5 семантический признак выявляемой словоформы задан первыми двумя символами кода признака (ТМ). Последние три символа кода замаскированы, т. е. игнорируются при сравнении. Результат 2 (Р2) имеется не в готовом виде, а формируется: переменной ТОР присваивается значение семантического признака из записи 1, выполненной данной сетью.

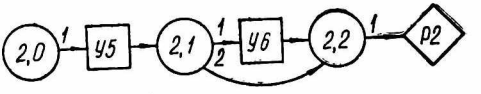

Рис. 2. Возможные пути последовательной аппроксимации

Сеть № 2 воспринимает такие фрагменты запросов: «(Дайте 1) плацкартный билет...», «...купейное место...», «...билет (в) общий (вагон)...», «...билет (на) купейное (место)...», «...место (в) купе...», «(Дайте 1) плацкарт...» и т. д.

Сети № 3, 4, ... имеют аналогичную структуру и принцип действия.

Особенность предлагаемой процедуры анализа фраз естественного языка заключается в том, что выявляется структура не одной или нескольких фраз в целом, а только структура сочетаний, выражающих существенные элементы значения текста. В процедуру анализа включены данные, относящиеся к морфологическому, синтаксическому и семантическому уровням языка. Можно анализировать запросы, состоящие из одной или нескольких фраз. В результате получаются элементы семантического представления текста.

Отдельные запросы с инвертированным порядком слов или другими особенностями построения фразы могут не восприниматься той или иной сетью. Распознающие возможности сети можно увеличить, расширив диапазон поиска в каждом условии $УК$ ($К$=1, 2, ...), для чего следует увеличить значение третьей компоненты признака Мвп. Однако это снижает быстродействие процедур анализа при безрезультатных исходах.

Анализ словосочетаний запроса продолжается до тех пор, пока не будут применены все сети, указанные в таблице. В результате получается множество переменных типа *SIT*, *TOP*, которым присваиваются значения из текста запроса. Возможен случай, когда значения получат переменные, не имеющие отношения к содержанию запроса. Это происходит, если некоторые сети ошибочно воспримут комбинации словоформ в запросе в качестве подходящих. Установление иерархии признаков, выделение релевантных переменных и их значений производится на следующем этапе анализа при семантическом анализе.

Список литературы: 1. *Попов Э. В.* Общение с ЭВМ на естественном языке.—М.: Наука, 1982.—360 с. 2. *Шенк Р.* Обработка концептуальной информации.—М.: Энергия, 1980.—358 с. 3. *Диалоговые* системы в АСУ/Под ред. Д. А. Поспелова—М.: Энергия, 1983.—216 с. 4. *Вудс У.* Сетевые грамматики для анализа естественных языков//Кибернет. сб. Нов. сер.—1976.—Вып. 13.—С. 120—158. 5. *Белоногов Г. Г., Кузнецов Б. А.* Языковые средства автоматизированных информационных систем.—М.: Наука, 1983.—288 с.

Поступила в редколлегию 14.11.85.