УДК 510.62

А. Ф. ОСЫКА, канд. техн. наук, *О. А. КРАВЕЦ*

К ВОПРОСУ О ФОРМИРОВАНИИ СЕМАНТИЧЕСКИХ ПРИЗНАКОВ СЛОВ

Одной из важных проблем, возникающих при создании систем машинного анализа и синтеза текстов на естественном языке, является формирование семантических признаков, которые сопоставляются словам в машинных словарях и используются при обработке связных текстов.

Роль семантического признака слова состоит в том, что он обозначает понятие и позволяет установить соответствие между этим понятием и конкретным объектом, признаком, действием или обстоятельством действия. Выбор семантических признаков зависит от предметной области, а также от структуры обрабатываемых текстов [1, 2]. В связи с этим в семантический признак слова заносится информация, которая обеспечивает распознавание осмысленности словосочетания или фразы, а также связь со знаниями о предметной области. Например, если речь идет о системе ввода запросов на естественном языке в базу данных, то семантические признаки слов определяются структурой самой базы данных, а также типами запросов, которые могут поступать в базу данных.

Структура семантического кода слова существенно зависит от используемого в данной системе семантического языка. В качестве такого языка наиболее часто применяются глубинные падежи, фреймы, язык исчисления предикатов первого порядка и т. д. [3, 4]. Семантический признак слова может иметь сложную структуру. Например, он может представлять собой семантическую сеть [5]. Но с целью ссылки на этот признак в словаре системы его целесообразно кодировать единым символом. Поэтому далее семантические признаки слов будут обозначаться отдельными символами.

В связи с ограниченностью машинных ресурсов в настоящее время автоматическая обработка смысла текста возможна лишь в заданной предметной области. При первоначальной настройке системы обработки текстов на некоторую предметную область разработчик лингвистического обеспечения снабжает словарные единицы системы выбранными семантическими признаками, по-

лученными на основе анализа объектов, их свойств и отношений в данной предметной области, а также анализа некоторого корпуса текстов, которые предположительно могут поступать от неподготовленных пользователей системы. В результате образуется множество M, которое имеет следующую структуру: $M = M_1 \cup M_2 \cup \ldots \cup M_n$, где $M_i (i = \overline{1, n})$ — множество слов и их семантических признаков, выполняющих одинаковую смысловую роль во фразе. Примерами таких ролей могут служить действие, его объект или субъект, а также более специальные роли, связанные, например, со структурой данных в базах данных. Причем M_i может рассматриваться как множество семантических кодов, соответствующих роли i в семантическом представлении текста, и как множество слов, которым сопоставлены эти коды в словаре.

Каждое M_i состоит из одного или нескольких подмножеств, объединяющих синонимические выражения. Всем таким выражениям ставится в соответствие единый семантический признак M_{ij}. Множество M слов и их признаков, которые первоначально включаются в словарь системы, можно представить в следующем виде:

$$\begin{cases} M_1 = M_{11} \cup M_{12} \cup \ldots \cup M_{1l_1}, \\ M_2 = M_{21} \cup M_{22} \cup \ldots \cup M_{2l_2}, \\ \quad \ldots \\ M_n = M_{n1} \cup M_{n2} \cup \ldots \cup M_{nl_n}, \end{cases}$$

где элементами подмножеств M_{ij} являются слова или словосочетания с общим признаком M_{ij}, который выражает лексическое значение элемента и его роль в предложении; а также, $\forall' j, k = \overline{1, l_i}$, $j \neq k M_{ij} \cap M_{ik} = \varnothing$.

Остановимся на втором условии, в соответствии с которым различные множества синонимичных выражений не могут иметь общих элементов. Для произвольных текстов это слишком сильное ограничение, что видно из последовательности слов: гений, талант, способный человек, неглупый человек, не дурак, посредственность, тупица, дебил. Крайние слова этой последовательности не только не является синонимами, но и имеют противоположное значение. Однако любые два соседних слова из приведенного перечня близки по смыслу. Таким образом, в общем случае отношение синонимии между парой слов не обладает свойством транзитивности. Это означает, что оно не является отношением эквивалентности и не может дать разбиения множеств M_i на непересекающиеся подмножества M_{ij}.

В приведенном примере лексические значения элементов последовательности отличаются степенью наличия некоторого качества (в данном случае — умственных способностей). Количество этого качества в различных объектах может изменяться непрерывно. С этим связаны нечеткость значений приведенных слов и словосочетаний, а также практически непрерывный переход значений одних слов в другие.

В технических системах отработки текстов, соотносимых с формальной моделью предметной области, ситуация несколько изменяется. Модель предметной области всегда имеет дискретный характер. Объекты, их свойства и отношения всегда четко определены, отличимы и противопоставлены друг другу. В связи с этим любые словесные формулировки, обозначающие некоторый объект или его свойство, имеют одинаковое значение в рамках данной предметной области и противопоставлены словесным обозначениям других объектов и свойств. Поэтому слова и словосочетания, описывающие один объект или свойство, образуют в машинных системах классы смысловой эквивалентности.

Изложенное выше не означает, что в технических системах не могут быть представлены и получить словесное выражение признаки объектов, принимающие непрерывное значение. Предположим, некоторая система располагает m объектами, каждый из которых обладает той или иной степенью следующих признаков: a_1 — умный, a_2 — молодой, a_3 — высокий, a_4 — тяжелый. Это означает, что в модели предметной области для указанных объектов заданы отношения частичного порядка по выделенным признакам: $a_{11} \leqslant a_{12} \leqslant \ldots \leqslant a_{1m}$, $a_{21} \leqslant a_{22} \leqslant \ldots \leqslant a_{2m}$, $a_{31} \leqslant a_{32} \leqslant \ldots a_{3m}$, $a_{41} \leqslant a_{42} \leqslant \leqslant \ldots \leqslant a_{4m}$.

Тогда элементы запросов типа «более высокий», «ниже», «умнее», «самый тяжелый» указывают на конкретное отношение между парой объектов или на граничный элемент этого отношения. Слова «гений», «мальчик», «великан» характеризуют положение элементов a_{1p}, a_{2g}, $a_{3k}(p, q, k=\overline{1, m})$ в соответствующих шкалах без привязки к другим элементам. Они сопоставляются объекту в словаре системы и характеризуют его однозначно. Если такое слово используется неподготовленным пользователем системы для обозначения другого объекта, необходима переформулировка запроса.

На вход системы обработки естественного языка могут поступать тексты, в которых содержатся слова, отсутствующие в машинном словаре. Для анализа таких слов им необходимо приписать семантические признаки и поместить в машинный словарь. Сведения о новом слове система может получить только от человека. Представляется целесообразным такой подход к приписыванию словам семантических признаков, при котором система осуществляет это в процессе диалога с пользователем, от которого поступил текст. При этом не предполагается, что пользователь знает структуру системы или используемые признаки слов. Вопросы системы должны быть сформулированы так, чтобы человек мог ответить утвердительно либо отрицательно, полагаясь только на свою языковую интуицию.

При формировании семантического признака может оказаться, что новое слово имеет одну из ранее выявленных семантических ролей M_i и является синонимом одного или нескольких слов, помещенных в словарь с признаком M_{ij}. Тогда и новому слову приписывается этот же признак. Возможен случай, когда при установ-

ленной роли M_i для нового слова не имеется синонимов в словаре. Тогда требуется организация еще одного класса эквивалентности с единственным элементом — новым словом, которому приписывается признак $M_ij(j>l_i)$. Рассмотрим случай, когда для нового семантически значимого слова среди имеющегося множества ролей не находится ни одной подходящей роли M_i. Под семантически значимым понимается слово во входном тексте, которому соответствует какой-нибудь элемент в словаре системы. Поэтому слова «пожалуйста», «побыстрее» и т. п. семантически значимыми не являются. Отсутствие подходящей роли для семантически значимого слова означало бы, что для некоторого вида выполняемых системой действий отсутствует соответствующее обозначение в виде семантической роли. Однако при введении новой функции системы или реквизита данных разработчик должен позаботиться о том, чтобы эта функция или реквизит были доступны пользователю. Поэтому появление новых возможностей исполнительного механизма должно сопровождаться обязательным формированием нового семантического класса $M_i(i>n)$, в котором должен содержаться хотя бы один класс эквивалентности, сформированный лицом, сопровождающим систему.

В данной работе исследуются некоторые вопросы, связанные с построением процедуры приписания семантических признаков новым словоформам в соответствии с изложенной выше постановкой. Возможный подход к формированию такой процедуры рассмотрим на примере системы, анализирующей команды пользователя, связанные с перемещением трех геометрических фигур — шара, конуса и куба.

Исходя из установленных правил, разрешается манипулировать выбранными объектами на бесконечной поверхности, расположенной горизонтально. На эту поверхность нанесена воображаемая система координат, ось абсцисс которой перпендикулярна линии взора пользователя, а ось ординат параллельна ей. Пользователь может подать любую команду, задающую относительное расположение двух фигур параллельно координатным осям. Не допускаются повороты фигур на любой угол относительно прямой, перпендикулярной рабочей поверхности. При каждой перестановке объекты должны касаться друг друга. Задача заключается в выборе такого набора семантических признаков, который обеспечивал бы успешное проведение анализа разнообразных команд пользователя.

Чтобы система могла успешно функционировать, требуется произвести ее предварительное обучение. Для этого до начала работы системы разработчик лингвистического обеспечения анализирует некоторый набор возможных команд пользователя, с тем чтобы сопоставить отдельным фрагментам этих команд (словам или словосочетаниям) признаки элементов действий, выполняемых системой. Предположим, этот набор состоит из четырех фраз: *Разместить куб левее шара, Шар поместить справа от конуса, Спереди от конуса поставить куб, Расположить по-*

зади куба; *конус*. Анализ этого множества команд позволяет лингвисту выделить четыре класса слов, причем элементы каждого класса присутствуют в любой команде.

Первый класс V включает в себя словоформы, обозначающие непосредственно вид действия. В данном случае элементами этого класса являются только глаголы в неопределенной форме. Однако, вообще говоря, здесь могут присутствовать и другие элементы естественного языка. Все глаголы в четырех анализируемых фразах являются синонимами с точки зрения действий, выполняемых в данном примере, поэтому они войдут в один класс эквивалентности V_1. Пусть v_{11} — разместить, v_{12} — поместить, v_{13} — поставить, v_{14} — расположить. Тогда $V=V_1=\{v_{11}, v_{12}, v_{13}, v_{14}\}$.

Второй класс F состоит из названий трех геометрических фигур, выступающих в качестве объектов выполняемых действий. Эти объекты образуют три класса эквивалентности, в каждом из которых находится по одному элементу. Обозначим f_{11} — куб, f_{21} — шар, f_{31} — конус. Тогда $F=F_1 \cup F_2 \cup F_3$, где $F_1=\{f_{11}\}$, $F_2=\{f_{21}\}$, $F_3=\{f_{31}\}$.

При описании предметной области человек использует некоторый набор отношений, которые связывают между собой объекты внешнего мира. Можно выделить количественные, признаковые, каузальные и другие отношения [3]. Для формализации действий с геометрическими фигурами представляют интерес отношения, указывающие на взаимное расположение объектов в некотором пространстве. Для их представления требуются математические категории направления, ширины, длины и т. д. В результате анализа предложенных команд можно выделить класс R слов, описывающих направление действия по отношению к некоторому объекту: $R=R_1 \cup R_2 \cup R_3 \cup R_4$, где $R_1=\{r_{11}\}=\{$левее$\}$, $R_2=\{r_{21}\}=\{$справа от$\}$, $R_3=\{r_{31}\}=\{$спереди от$\}$, $R_4=\{r_{41}\}=\{$позади$\}$.

Следующий класс P формируется из так называемых объектов-ориентиров — геометрических фигур, относительно которых перемещаются объекты действия: $P=P_1 \cup P_2 \cup P_3$, где $P_1=\{p_{11}\}=\{$шар$\}$, $P_2=\{p_{21}\}=\{$конус$\}$, $P_3=\{p_{31}\}=\{$куб$\}$.

Таким образом, множество M в данном случае состоит из четырех элементов, где $M_1=V$, $M_2=F$, $M_3=R$, $M_4=P$.

Наличие в команде элементов каждого из четырех установленных выше классов обязательно, так как без любого из них она теряет смысл. В рамках каждого класса эти элементы взаимозаменяемы. Однако замена любого слова команды словом или словосочетанием из другого множества, как правило, приводит к ее обессмысливанию. Например, команда *Поместить левее позади шара* не может быть выполнена, так как она противоречит нормам русского языка.

Отметим, что сущность замен элементов внутри множеств F, R, P и V различна. Поскольку для данного примера глаголы класса V являются синонимами, их взаимные замены не влияют на

суть действия, которое необходимо выполнить. При смене любого из элементов множеств F, R, P значение команды меняется.

Рассмотрим, как работает система, получив от пользователя команду поместить некоторую геометрическую фигуру на определенное место поверхности по отношению к другой фигуре. Вначале система выясняет, все ли слова команды она в состоянии идентифицировать. Отрицательный результат дает ей повод обратиться к пользователю с предложением объяснить одно или несколько незнакомых слов полученной фразы.

Предложим, на вход системы поступила команда α: разместить куб правее конуса. Все ее слова будут идентифицированы, кроме словоформы «правее». Поскольку в команде не выявлен только элемент класса R, неизвестная словоформа должна принадлежать этому классу. В противном случае команда не может быть выполнена. Первым шагом при проверке принадлежности незнакомого слова множеству R является построение диагностической фразы, в которой все идентифицированные слова команды сохраняют свой порядок следования, а на месте незнакомой словоформы оставляется свободное место. Эта диагностическая фраза, обозначаемая d_3, выглядит следующим образом: *Разместить куб* |__| конуса.

После этого система выбирает из множества R какой-нибудь элемент (например, первый) и помещает его на свободное место диагностической фразы d_3. В результате получается выражение w_{31}: *Разместить куб левее конуса.*

Теперь система обращается к пользователю с вопросом о том, имеет ли w_{31} сходный смысл с исходной командой α. Предварительно человеку разъясняется, что нужно подразумевать под выражением «сходный смысл». Для этого система выдает, например, следующее сообщение s_3. П о л а г а е м, ч т о ф р а з ы:

Поставить конус левее шара и *поставить конус позади шара*

и м е ю т с х о д н ы й с м ы с л.

В этом сообщении элементы множеств V, F и P одинаковы для двух фраз, а элементы множества R принадлежат разным классам эквивалентности. После выведения s_3 пользователю предъявляется вопрос: И м е ю т л и с х о д н ы й с м ы с л ф р а з ы:

Разместить куб левее конуса и Разместить куб правее конуса?

Получив положительный ответ, система имеет основание занести словоформу «правее» в класс R. Для введения указанной словоформы в конкретный класс эквивалентности необходимо вновь применить диагностическую фразу d_3. При подстановке на ее свободное место представителей классов эквивалентности $R_j (j=\overline{1, l_3})$ получаются фразы $w_3 j$, которые система предъявляет пользователю для сравнения с командой α. Пользователь должен ответить,

обладают ли $w_3 j$ и α одинаковым смыслом. Элементы множества R перебираются в порядке возрастания номеров классов эквивалентности. Поэтому вопрос, на который человек ответит утвердительно, поступит вторым. Он формулируется так: имеют ли одинаковый смысл фразы:

Разместить куб справа от конуса и *Разместить куб правее конуса?*

Каждое новое слово после установления его значения вносится в список элементов соответствующего класса. При этом оно либо входит в класс эквивалентности своего синонима, либо образует свое подмножество с номером $\max(l_i)+1$. В данном примере во множестве R нашелся элемент r_{21}, принадлежащий классу R_2, при подстановке которого в диагностическую фразу она совпала по смыслу с поступившей командой. По этой причине не идентифицированное ранее слово «правее» становится элементом r_{22} уже существующего класса эквивалентности R_2. Множество R теперь приобретает новый вид

$$R=\{r_{11}\}\cup\{r_{21},\ r_{22}\}\cup\{r_{31}\}\cup\{r_{41}\}.$$

Обобщим приведенную выше процедуру на анализ любой команды β. Опишем алгоритм функционирования системы, когда в команде не идентифицировано одно из слов x, которое может быть элементом множеств $M_1, M_2, \ldots, M_n$.

1. Установление принадлежности x одному из $M_i(i=\overline{1,n})$.
2. Формирование диагностической фразы d_i.
3. Образование w_{i1} путем подстановки первого элемента множества M_i в d_i.
4. Пояснение пользователю значения выражения «сходный смысл».

Вывод с этой целью сообщения s_i.

5. Вопрос: *Имеют ли сходный смысл* w_{i1} и β?

Ответ: Нет — прекращение работы. Команда не выполнима.

Ответ: Да — переход к п. 6.

6. $z=1$.
7. Вопрос: Имеют ли одинаковый смысл w_{iz} и β?

Ответ: Да — x становится элементом M_{iz}.

Ответ: Нет — переход к п. 8.

8. $z=z+1$.
9. Если $z\neq l_i+1$, то образование w_{iz} и возврат к п. 7; в противном случае — формирование нового класса эквивалентности $M_{iz}=\{x\}$.

Обозначим приведенный алгоритм через A. Проследим этапы его работы при измененных условиях функционирования системы. Пусть теперь манипуляции с тремя геометрическими фигурами разрешено производить не на поверхности, а в пространстве. Другими словами, к четырем существующим классам пространственных отношений добавляется пятый, который описывает размещение некоторого объекта сверху другого.

Пусть на вход поступила команда, расположить шар сверху куба. Система выполняет следующие действия:

1. Определяет, что в её словаре содержатся все слова команды, кроме слова из множества R.

2. Формирует диагностическую фразу d_3: *Расположить шар* |__| *куба.*

3. Образует w_{31}: *Расположить шар левее куба.*

4. Выводит сообщение s_3: Полагаем, что фразы *Поставить конус левее шара* и *Поставить конус позади шара* имеют сходный смысл.

5. Задает вопрос пользователю: Имеют ли сходный смысл фразы *Расположить шар левее куба* и *Расположить шар сверху куба?* Получает ответ: Да.

6. Имеют ли одинаковый смысл фразы *Расположить шар левее куба* и *Расположить шар сверху куба?* — Нет.

7. Образует w_{32}: *Расположить шар справа от куба.*

8. Имеют ли одинаковый смысл фразы *Расположить шар справа от куба* и *Расположить шар сверху куба?* — Нет.

9. Образует w_{33}: *Расположить шар спереди от куба.*

10. Имеют ли одинаковый смысл фразы *Расположить шар спереди от куба* и *Расположить шар сверху куба?* — Нет.

11. Образует w_{34}: *Расположить шар позади куба.*

12. Имеют ли одинаковый смысл фразы *Расположить шар позади куба* и *Расположить шар сверху куба?* — Нет.

13. Формирует новый класс эквивалентности $M_{35}=R_5=\{r_{51}\}=$ $=\{$сверху$\}$.

Алгоритм A применим и в том случае, когда в команде не идентифицированы два или более слов. Он используется столько раз, сколько слов не удалось понять системе. Для сокращения вариантов подстановок в диагностические фразы слова следует отбирать с учетом их морфологических признаков.

Предположим теперь, что при манипулировании геометрическими фигурами разрешается поворачивать их на 180° вокруг горизонтальной оси, причем переворачивать можно только перемещаемые в соответствии с прежними условиями объекты.

Появление новых возможностей исполнительного механизма системы сопровождается образованием нового семантического класса M_5. В нем должен содержаться хотя бы один класс эквивалентности, о формировании которого обязан позаботиться разработчик лингвистического обеспечения системы. Проанализировав одну из возможных команд пользователя *Поместить повернутый на 180° шар левее конуса,* можно выделить элемент t_{11} — *повернутый на 180°* и приписать его классу эквивалентности T_1. С учетом появления нового класса $T=M_5$ изменилась структура множества M. Теперь его можно представить в следующем виде: $M=$ $=V \cup F \cup R \cup P \cup T$.

Дополнительные возможности манипулирования объектами делают неопределенной фреймовую структуру команд пользователя. Если в первых двух случаях для представления команды было достаточно одного фрейма $\Phi_1=\{V, F, R, P\}$, то теперь появилась возможность выполнения новых действий и ввода новых типов

команд. Из-за этого затрудняется работа с алгоритмом A пополнения уже существующих и образования дополнительных классов эквивалентности. Это происходит из-за невозможности образования диагностических фраз единого типа для всевозможных команд пользователя.

При измененных условиях работы системы команды пользователя могут быть представлены тремя фреймами: Φ_1, $\Phi_2=\{V, F, R, P, T\}$ и $\Phi_3=\{T, F\}$. Фрейм Φ_2 реализуется в команде *Разместить перевернутый куб справа от конуса, а* Φ_3 — в команде *Перевернуть шар*. Предположим, система получает некоторую команду и идентифицирует такие ее слова, по которым однозначно определяется ее фреймовая структура. Тогда приписание семантических признаков незнакомым словоформам сводится к выполнению алгоритма A. В противном случае пользователь получает сообщение о том, что его команда не может быть выполнена, так как система не обладает достаточным объемом знаний для проведения ее анализа.

В заключение перечислим действия, выполняемые системой для анализа некоторой команды пользователя.

1. Идентификация словоформ команд с целью установления ее фреймовой структуры.

2. Если фреймовая структура определена однозначно, переход к п. 4; в противном случае — к п. 3.

3. Применение дополнительной процедуры снятия фреймовой неоднозначности, разработка которой не входит в рамки данного исследования.

4. Установление числа y незнакомых слов команды.

5. Если $y=1$, переход к п. 7.

6. Проведение морфологического анализа y незнакомых словоформ с целью увеличить вероятность их правильной подстановки в диагностические фразы.

7. Применение алгоритма A y раз с последовательной подстановкой каждого незнакомого слова в диагностические фразы.

Процедура формирования и пополнения классов семантических признаков может быть использована в машинных системах обработки естественного языка, основанных на дискретной модели предметной области, включающей значительное число объектов и отношений между ними.

Список литературы: 1. *Апресян Ю. Д.* Лексическая семантика. М., 1974. 324 с. 2. *Диалоговые* системы в АСУ//Под ред. Поспелова Д. А. М., 1983. 204 с. 3. *Поспелов Д. А.* Ситуационное управление: теория и практика. М., 1986. 288 с. 4. *Попов Э. В.* Общение с ЭВМ на естественном языке. М., 1982. 360 с. 5. *Шенк Р.* Обработка концептуальной информации. М., 1980. 358 с.

Поступила в редколлегию 14.02.90