УДК 004.048

Б. Н. СУДАКОВ, канд. техн. наук, профессор, НТУ «ХПИ»;
М. Ю. ЯРМАК, магистр, НТУ «ХПИ»

ФОРМАЛЬНОЕ ПРЕДСТАВЛЕНИЕ СЕМАНТИКИ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ В ЭКСПЕРТНЫХ СИСТЕМАХ

В статье рассматривается проблема семантики в естественно-языковых текстах и проблемы с ее формальным представлением. Статья предоставляет обзор моделей, описывающих семантику естественно-языковых текстов. Определена и обоснована необходимость создания новых моделей внутреннего языка с использованием аппарата формальных грамматик. Предложено расширение моделей формальной грамматики Хомского семантическими продукционными правилами.

Ключевые слова: семантика, экспертные системы, внутренний язык, формальная грамматика, модель «СМЫСЛ - ТЕКСТ», модель языка.

Анализ последних исследований и литературы. Автоматизация процессов принятия решений (ПР) в различных областях человеческой деятельности на основе развития новых информационных технологий находит в настоящее время широкое применение. Анализ применения интеллектуальных систем (ИС) различного типа показал, что эффективность от их использования в современных условиях зависит от многих системотехнических факторов: возможности описания логико-лингвистических задач и накопления опыта, наличия механизмов логического вывода и естественно-языкового интерфейса (ЕЯИ), удобства пользования, полноты базы знаний и достоверности результатов и др.

Одним из элементов экспертных систем (ЭС) для ПР является подсистема взаимодействия с пользователем. Основу взаимодействия составляют языковые средства, поскольку только с помощью языка (формального или естественного) можно достичь определенных целей в процессе общения коммуникантов.

К настоящему времени вызывают значительные трудности вопросы формализации семантики и прагматики естественного языка (ЕЯ). К тому же возникает вопрос о размытости границ семантики и прагматики, что еще больше усложняет процесс формализации ЕЯ. Выходом из создавшегося положения является как ограничение ЕЯ рамками предметной области, так и наложение ограничений по структуре предложений.

Также, среди всех существующих моделей языков взаимодействия выделяются модели семантической ориентации, которые предлагают осуществлять процесс трансляции под управлением семантики, а синтаксический компонент привлекать лишь в случае появления неоднозначностей.

Цель статьи. Рассмотреть и проанализировать модели семантической

ориентации и построить модель внутреннего языка с использованием аппарата формальных грамматик.

Основная часть. Существует 3 модели языков взаимодействия:

1) Элементарные теоретико-множественные модели, в которых текст представляется как неупорядоченное множество лексических единиц

2) Модели синтаксической ориентации. Характеризуются тем, что в них проводится сложный синтаксический анализ, базирующийся на какой-либо формальной модели описания синтаксиса языка.

3) Модели семантической ориентации предлагают осуществлять процесс трансляции под управлением семантики, а синтаксический компонент привлекать лишь в случае появления неоднозначностей.

Подробно проанализируем интересующую нас модель семантической ориентации.

Одной из первых моделей, отражающих состояние лексической семантики, являлась модель компонентного анализа. Данная модель основана на том, что семантика языков адекватно выражена в виде конечного неструктурированного набора семантических множителей (атомов смысла), разбивающих слова на разные семантические группы. Значение каждого слова представляется как множество атомов смысла, которые затем объединяются в смысловые словосочетания. Использование данной модели требует применения чрезвычайно чувствительных дифференцированных признаков. Еще большие трудности возникают при попытке выразить смысл целого предложения. Отмеченные недостатки не означают, что компонентный анализ надо полностью отвергнуть. Возможно, он имеет право на ограниченное использование, например, при выделении атомов смысла. Однако этот подход не решает все проблемы семантики.

Придерживаясь основной идеи компонентного анализа и аргументной структуры предиката, Филмор предлагает указывать не только число аргументов, но и описывать их семантическое содержание. Данная модель получила название модели семантических падежей. Она имеет много общего с используемым в отечественной литературе понятием «модель управления» (МУ). МУ является описанием семантико-синтаксических свойств слова. Имея дополнительные возможности по выявлению атомов смысла, данная модель не является завершенной и не позволяет описывать всю семантическую структуру фразы, но в тоже время используется в более сложных моделях, таких как СМЫСЛ-ТЕКСТ.

Типичными представителями моделей СМЫСЛ-ТЕКСТ являются модели концептуальных зависимостей, семантик предпочтения и непосредственно модель «СМЫСЛ-ТЕКСТ» (Жолковского). Модели СМЫСЛ-ТЕКСТ различаются между собой прежде всего тем, как представляется семантика языка. Основой модели концептуальных зависимостей является квазиграф: кроме бинарных отношений в нем есть тернарные (типа X переходит от Y к Z) и кватернарные (типа «состояние X изменилось с Z_1 на Z_2 по параметру У»),

дуги квазиграфа (в отличие от графа) связывают не только вершины, но и другие дуги.

Модель семантик предпочтения для выражения сущностей выделяет: семантические формулы – для выражения смысла слова; «образец» для представления сообщения; правила следования – для выражения правил семантической совместимости. Описанная структура представляет собой заключенные в скобки семантические элементы.

Модель «СМЫСЛ-ТЕКСТ» (Жолковского) для представления смысла использует два компонента: семантический граф и сведения о коммуникативной организации смысла. Семантический граф – это связный ориентированный граф. Вершины графа помечены атомами смысла, дуги изображают связь предиката с его аргументами. Всем моделям СМЫСЛ-ТЕКСТ присущ ряд недостатков.

Ситуации рассматриваются на одном уровне детальности, а это не позволяет описывать сложные события через более простые термины. Для систем реальной степени сложности необходимо варьировать уровень детальности в зависимости от решаемых задач.

Модели СМЫСЛ-ТЕКСТ по своей сущности являются моделью языка, что приводит к нечеткому выделению языковых средств для описания ПО. В таких моделях информация об окружающем мире сводится только к словам языка, а не к событиям и процессам, имеющим место при моделировании предметной области.

Большинство моделей СМЫСЛ-ТЕКСТ ориентированы только на анализ языка, а те модели, которые используют синтез (например, система ПОЭТ) очень громоздки, так как процедуры анализа и синтеза независимы друг от друга, что значительно усложняет модели.

Таким образом, на основании вышеизложенного можно сформулировать вывод, что для построения модели внутреннего языка наиболее целесообразно использовать аппарат формальных грамматик, по возможности расширив их правилами, позволяющими учитывать семантику предметной области.

Для представления в ЭВМ формул исчисления необходима формальная модель внутренне-языковой (ВЯ) системы. Под моделью языка (МЯ) понимается следующая система

$$\text{МЯ} < C, P, A >, \tag{1}$$

где C – множество базовых элементов языка или словарь; P – множество правил, позволяющих из С строить синтаксически правильные конструкции языка; A – множество априорно- истинных конструкций, называемых аксиомами.

Рассмотрим структуру словаря. Исходя из того, что основными единицами языка, которые необходимы для построения более сложных конструкций, являются словоформы профессионального языка пользователей, каждый элемент C можно охарактеризовать следующей информацией

$$S_i\ (tS_i, sint_i, sem_i), \tag{2}$$

где S – i-я словоформа словаря C; *sint* – синтаксическая информация, характеризующая S_i; *sem* – совокупность семантической информации, которая может быть приписана S_i (признак, характеристика, класс и т.д.) в зависимости от классификации объектов; tS_i – тип словоформы, относящий S_i к терминальным T либо нетерминальным N символам словаря.

Таким образом, словарь разбивается на два непересекающихся множества V_N и V_m ($V_N \cap У_m = Ø$), а $C = V_N \cup V_m$. В соответствии с теорией формальных грамматик из символов С строятся цепочки типа $S_1\ S_2\ S_э\ ...\ S_n$, которые считаются ориентированными слева направо. Если цепочка является пустой не содержит ни одного символа), то она обозначается через $C^\circ = Λ$.

Нетрудно показать, что множество всех возможных цепочек (его еще называют замыканием C) алфавита определяется как

$$C^* = U_{n=0}^{\infty} = C^n . \qquad (3)$$

Для удобства определим также множество непустых цепочек над C следующим образом

$$C^+ = C^* \setminus \{\Lambda\} = U_{n=1}^{\infty} C^n . \qquad (4)$$

Основная операция, осуществляемая над строками, называется конкатенацией. Формально она может быть определена как бинарная операция • на C^* следующим образом: • :(α, β) → αβ, где α и β – произвольные цепочки. Необходимо заметить, что по отношению к операции• C^* является моноидом, а $C+$ – полугруппа. Любое множество цепочек $L \in C$ называется формальным языком и описывается с помощью ФГ $G(L)$, основы которых были заложены Хомским. В соответствии с изложенным ФГ для описания всех структур объектов может быть задана в следующем виде

$$G(L) = (V_T, V_N, P, A). \qquad (5)$$

В исследованиях по методам анализа цепочек языка предполагают две стратегии грамматического разбора: нисходящая и восходящая.

Работа нисходящего распознавателя теоретически основывается на идее использования порождающей грамматики при генерации всех возможных цепочек языка, пока не будет порождена цепочка соответствующая входной. Для этого необходимо предусмотреть проверку альтернатив и способов возврата из тупиков. Такая ситуация возникает, когда имеются продукции с одинаковыми левыми частями вида $A \to \alpha$, $A \to \beta$, где $A \in V_N$.

Стратегия восходящего разбора состоит в том, что во входной цепочке ищутся одинаковые правые части продукции с одинаковыми левыми частями. Процесс повторяется до тех пор, пока не получится начальный символ грамматики. Основной проблемой восходящего разбора является проблема выбора альтернатив, такая ситуация возникает, когда есть продукции с одинаковыми правыми частями вида $A \to \alpha$, $B \to \beta$, где $A, B \in V_N$.

Исходя из этого выходит, что грамматики Хомского не учитывают семантику и поэтому бывает трудно выйти из тупиковых ситуаций при разборе входного сообщения. Поэтому предлагается использовать расширение моде-

лей Хомского, в которых учитывать семантические признаки *sem*, применяя семантические продукционные правила (СПП) вида

$$S_k\,(sem_k),\ P_z \rightarrow S_j\ semj),\ P_m. \qquad (6)$$

Данное СПП означает, что если S_k с соответствующим семантическим признаком присутствует в правиле переписывания с номером z, то в правиле с номером m должна присутствовать словоформа с семантическим признаком sem_j. Это позволяет осуществлять проверку на семантическую корректность цепочек внутреннего языка. Чтобы осуществлять выбор альтернатив продукционное правило (6) интерпретируется следующим образом. Если S_k с соответствующим семантическим признаком присутствует в правиле переписывания с номером z, то при разборе входной цепочки языка выбирается правило с номером m и те S_j, которые имеют семантический признак sem_j.

С учетом изложенного ФГ будет иметь следующий вид:

$$G(L) = (V_T, V_N, P, SP, A), \qquad (7)$$

где SP – семантические правила, а V_m, V_N – терминальный и нетерминальный словари соответственно.

Среди единиц терминального словаря можно выделить элементы, которые являются объектами определенных категорий рассматриваемой предметной области. Это согласуется с представлением ПО у человека, когда он во внешнем мире вычленяет конечный набор объектов и отображений. Отображения (отношения) связывают между собой объекты внешнего мира. Таким образом, словоформам терминального словаря может быть приписана семантическая информация в соответствии со структурой предметной области.

Необходимо заметить, что заполнение некоторых элементов формул возлагается полностью на пользователей. Это требует дополнительных знаний о семантике предметной области. Например, если объект выполняет роль «являться признаком времени» или «признаком пространства», то в качестве значения признака должны стоять словоформы, характеризующие время или пространство. Это можно учесть с помощью все тех же семантических правил.

Выводы. Использование модели компонентного анализа требует применения чрезвычайно чувствительных дифференцированных признаков. Еще большие трудности возникают при попытке выразить смысл целого предложения. Таким образом, его использование достаточно ограничено.

Это дало толчок к разработке модели семантических падежей. Имея дополнительные возможности по выявлению атомов смысла, данная модель не является завершенной и не позволяет описывать всю семантическую структуру фразы, но в то же время используется в более сложных моделях, таких как СМЫСЛ-ТЕКСТ.

Однако для построения модели внутреннего языка наиболее целесообразно использовать аппарат формальных грамматик, по возможности расширив их правилами, позволяющими учитывать семантику предметной области.

В результате проведенного анализа можно заключить, что наиболее целесообразным при разработке ЭС является использование ограниченного ЕЯ поль-

зователей и внутреннего языка системы, на котором представляются знания. Кроме того, указанные языки должны быть семантически расширяемы и, по возможности, обеспечивать простоту трансляции с одного на другой. Должно быть обеспечено также требуемое качество лингвистического обеспечения, характеризующее естественность взаимодействия пользователя с системой, необходимую глубину проникновения в смысл, возможность описания предметной области с учетом неоднозначности и неопределенности.

Список литературы: 1. *Девятков В. В*. Системы искусственного интеллекта / Гл. ред. *И. Б. Федоров*. – М.: Изд-во МГТУ им. *Н. Э. Баумана*, 2001. – 352 с. **2.** *Люгер Дж. Ф.* Искусственный интеллект: стратегии и методы решения сложных проблем = Artificial Intelligence: Structures and Strategies for Complex Problem Solving / Под ред. *Н. Н. Куссуль*. – М.: Вильямс, 2005. – 864 с. **3.** *Джозеф Джарратано, Гари Райли* Экспертные системы: принципы разработки и программирование : Пер. с англ. – М. : Издательский дом «Вильямс», 2006. – 1152 стр. **4.** *Питер Джексон*. Введение в экспертные системы = Introduction to Expert Systems. – М.: Вильямс, 2001. 624 с. **5.** *Рыбина Г. В*. Основы построения интеллектуальных систем. – М.: Финансы и статистика; ИНФРА-М, 2010. – 432 с.

Поступила в редколлегию 11.10.2012

УДК 004.048

Формальное представление семантики естественно-языковых текстов в экспертных системах / Б. Н. Судаков, М. Ю. Ярмак // Вісник НТУ «ХПІ». Серія: Техніка та електрофізика високих напруг. – Х.: НТУ «ХПІ», 2012. – № 52 (958). – С. 184-190. – Бібліогр.: 4 назв.

У статті розглядається проблема семантики в природно-мовних текстах і проблеми з її формальним представленням. Стаття надає огляд моделей, що описують семантику природномовних текстів. Визначена й обгрунтована необхідність створення нових моделей внутрішнього мови з використанням апарату формальних граматик. Запропоновано розширення моделей формальної граматики Хомського семантичними продукційними правилами.

Ключові слова: семантика, експертні системи, внутрішня мова, формальна граматика, модель «СЕНС-ТЕКСТ»", модель мови.

The article discusses a problem of semantics of natural language texts, and the problems with its formal submission. The article provides an overview of models describing the semantics of natural language text. Define and justify the need for new models of the internal language using the apparatus of formal grammars. It was proposed extension of models of formal grammar Chomsky by semantic production rules.

Keywords: semantics, expert systems, internal language, a formal grammar, the model «SENSE – TEXT», language model.