Некоторые подходы к проблеме создания программы приобретения знаний на основе анализа текстов задач на естественном языке

Кострюков С.

В статье рассматриваются следующие вопросы: способы приобретения знаний посредством программной системы; возможные структуры программ приобретения знаний; инвариантное ядро программ приобретения знаний; сравнение с экспертными системами; способы использования программ приобретения знаний; уровень знаний о приобретаемых знаниях в программах приобретения знаний.

При решениии задач конкретной предметной области в интеллектуальной системе автоматизированного программирования используются различные виды семантических сетей. Дуга семантической сети нагружена отношениями между объектами, процессами, явлениями, которые представляются вершинами. Для примера рассмотрим следующую задачу : «В вершинах A_1, B, D_1 куба $ABCDA_1B_1C_1D_1$, ребро которого равно a, помещены точечные заряды q. Выразить результирующую напряженность создаваемого ими электрического поля в точках A и C_1 через вектор AC_1. Найти абсолютную величину напряженности в точках C и B_1, в центре грани $A_1B_1C_1D_1$ и в центре куба. Напряженность определяется по следующему правилу. Если в точке O находится точечный заряд q, то напряженность **E**, создаваемого им электрического поля в точке M выражается формулой $\mathbf{E} = (k*q)/OM^2*\mathbf{OM}$, где k зависит от выбора системы единиц».

По сформулированному тексту в создаваемой системе «Интеллектуальная система автоматизированного программирования» создаются семантические сети задачи первого (ЗдчСемС1) и второго (ЗдчСемС2) уровней. В семантической сети первого уровня выделяется словосочетания и образуется семантическая сеть, учитывающая смыслы, изложенные в толковых словарях. Семантическая сеть второго уровня привязывается к предметной области задачи, принимая во внимание классификационные (КлСемС), функциональные (ФСемС), понятийные (ПСемС) семантические сети. При стыковке КлСемС, ФСемС, ПСемС и ЗдчСемС1 создается ЗдчСемС2. Семантические сети в данном докладе представляются в виде списков, структура которых такова: (Порядковый номер списка(Начальная вершина(Возможен неопределенный вложенный список дополнительной информации))(Конечная вершина(Возможен неопределенный вложенный список дополнительной информации))(Ссыдка)Отношение[или связь])

Допустим, что в семантических сетях имеется следующая информация. В КлСемС содержится :

(1 ((Геометрический объект)(Точка((Заглавная буква без индекса или с индексом)
(Точка – вид геометрического объекта))Вид)
(2 (Геометрический объект)((Плоский геометрический объект)())Вид)
(3(Геометрический объект)((Пространственный геометрический объект)())Вид)
(4(Геометрический объект)((Линия)())Вид)
(5(Плоский геометрический объект)((Линия)())Вид)
(6(Плоский геометрический объект)((Плоскость)(Три заглавные буквы без индекса или с индексом))Вид)
(7(Плоский геометрический объект)((Фигура)())Вид)

В ФСемС имеется:

(1(Значение 1(Слагаемое))(Сложение(Операция +)(Символ «+» или «add»)) входное данное)
(2(Значение 2(Слагаемое))(Сложение(Операция +)(Символ «+» или «add»))входное данное)
(3(Сложение(Операция +))(Сумма(Результат)())выходное данное)
(4(Значение 3 (Уменьшаемое)) (Вычитание(Операция -)(Символ «-» или «sub»))входное данное)
(5(Значение 4(Вычитаемое))(Вычитание(Операция -)(Символ «-» или «sub»))входное данное)
(6(Вычитание(Операция -))(Разность(Результат)(выходное данное))

В ПСемС имеется:

(1(Вершина куба(Определяемое понятие))(Определение вершины куба(Определение: Пересечение трех смежных граней)(Заглавные буквы алфавита без индекса или с ндексом))Есть)
(2(Грань куба(Определяемое понятие))(Определение грани куба(Определение: Конечная часть ограничивающей плоскости куба, ограниченная попарно смежными ребрами, лежащими в этой же плоскости)())Есть)
(3(Точечный заряд(Определяемое понятие))(Определение понятия(Определение: Первичное понятие. Имеет электрический заряд)()Есть)

В ЗдчСемС1 имеется:

(1(Точечные заряды(Словосочетание))(q(Слово))Какие)
(2(Точечные заряды(Словосочетание))(В вершинах(
Словосочетание))Помещены)
(3(В вершинах(Словосочетание))(A_1, B, D(Словосочетание))Каких)
(4(В вершинах(Словосочетание))(куба $ABCDA_1B_1C_1D_1$(Словосочетание))Чего)
(5(Ребро(Слово))(куба $ABCDA_1B_1C_1D_1$(Словосочетание))Чего)
(6(Ребро(Слово))(a(Слово))Равно)
(7(Результирующую напряженность(Словосочетание))(Через вектор AC_1(Словосочетание) Выразить)
(8(Результирующую напряженность(Словосочетание))(Электрического поля (Словосочетание))Чего)
(9(Электрического поля(Словосочетание))(Создаваемого точечными зарядами q (Словосочетание))Какого)
(10(Создаваемого точечными зарядами q(Словосочетание))(В точках A и C_1 (Словосочетание))Где)
(11(Абсолютную величину результирующей напряженности(Словосочетание))(В точках C и B_1 (Словосочетание))Найти)
(12(Абсолютную величину результирующей напряженности(Словосочетание))(В центре грани $A_1B_1C_1D_1$(Словосочетание))Найти)
(13(Абсолютную величину результирующей напряженности(Словосочетание))(В центре куба(Словосочетание))Найти)
(14(Напряженность(Словосочетание))(По следующему правилу(Словосочетание)) Определяется)
(15(Если в точке O(Словосочетание))(Точечный заряд q(Словосочетание))Находится)
(16(То напряженность **E**(Словосочетание))(Формулой $\mathbf{E} = (k*q)/OM^3*\mathbf{OM}$ (Словосочетание))Выражается)
(17(Кэффициент k(Словосочетание))(От выбора системы еиниц(Словосочетание))Зависит)

В приведенных списочных представлениях описано содержание семантических сетей, существующее до разбора данной текущей задачи и приведена семантическая сеть первого уровня задачи.

В классификационной семантической сети отсутствует информация о словосочетаниях: «точечные заряды», «результирующая напряженность», «электрическое поле», «абсолютная величина», «центр грани», «центр куба», «система единиц»; о словах: «куб», «вершина», «вектор», «грань», «формула», «коэффициент», «выбор», «единица». Отсутствие обнаруживается при сопоставлении КлСемС и ЗдчСемС1. Аналогично при сопоставлении ПСемС и ЗдчСемС1 обнаруживается отсутствие понятий «точечный заряд», «вершина», «куб» и т.д.

При сопоставлении классификационных семантических полей происходит выделение понятий и слов с помощью АСемС в каждой вершине и соответственно в каждой характеристике, если она присутствует, на основании морфологической, синтаксической и семантической информаций, хранящихся для каждой вершины семантических сетей. Анализатор семантических сетей (АСемС) может вызываться из любого программного приложения с указанием семантической сети и режима работы. Примеры работы АСемС приведены в докладе «К вопросу о создании интеллектуального интерфейса с организацией обработки запроса на естественном языке». После работы АСемС в ПСемС появляются неопределяемые понятия, которые определяются с помощью ИИн(Интеллектуальный Интерфейс) и ППрЗн(Программа Приобретения Знаний). С помощью компоненты КСемС (Корректировка Семантической Сети) можно удалить, изменить, вставить информацию в семантическую сеть. С помощью ССемС (Сопоставителя Семантических Сетей) для двух указанных сетей и определения базовой сети выделяются, в зависимости от указанного режима работы, понятия, отношения, характеристики и дополняются недостающими элементами в базовой семантической сети. Примеры работы КСемС приведены в докладе «К вопросу о создании интеллектуального интерфейса с организацией обработки запроса на естественном языке».

Приступать к решению задачи можно только в том случае, когда известна вся информация об условиях, ограничениях, требованиях, входных данных, целях задачи, что представлено в семантической сети задачи.

В ЗдчСемС1 приведены словосочетания, которые не все могут быть понятиями. Значит для этого словосочетания надо разбить на более мелкие словосочетания или слова. Процесс приобретения знаний состоит в выполнении пяти основных функций: анализа словосочетаний и шаблонов задач, сопоставлении, нахождении источника знаний, создании шаблонов задач, записи полученной информации в нужное место. При сопоставлении один объект должен являться базовым, а другой входным, при этом должно быть указано, что сопоставляется: понятия, характеристики, отношения и т.д. Операции сопоставления в математике соответствует операция соответствия. Семантические сети, являющиеся базами знаний, всегда являются базовыми по отношению к семантическим сетям задач. Кратко алгоритм сопоставления состоит в следующем. Выбирается очередная начальная вершина в базовой семантической сети и анализируется словосочетание, связанное с этой вершиной на предмет соответствия его указанному виду сопоставления. После выделения элемента сопоставления происходит контекстный поиск во входной семантической сети(ВхСемС) с целью выявления повторяемости. В случае индивидуальной встречаемости выделенного элемента сопоставления во ВхСемС делается вывод о том, что элемент сопоставления определен в ВыхСемС. Если элемент сопоставления встречается в только в контексте, то это означает, что он используется в определениях, но сам не определен. Если элемент сопоставления нигде не встречается, то это означает, что он не определен и сам не используется для определения. В последних двух случаях нужно находить источник знания. Сначала надо обратиться к функциональной семантической сети, нет ли там функционального

определения элемента сопоставления. При наличии элемента сопоставления в функциональной сети посредством ИИн организуется диалог для доопределения элемента сопоставления. При отсутствии элемента сопоставления рассматривается КлСемС с целью организации диалога с помощью ИИн и найденной информации в сети. В этом случае ППрЗн работает с КлСемС следующим образом. Выбирается очередная начальная вершина с текстом «ТТТ…ТТТ». Допустим элементу сопоставления соответствует текст «ССС…ССС». В этом случае ППрЗн формирует вопрос типа ««ССС..ССС» является видом «ТТТ..ТТТ» ?Ответ(Да,Нет)». В случае положительного ответа формируется строка в КлСемС, в которой начальной вершине сответствует текст «ТТТ..ТТТ», а конечной вершине – текст «ССС…ССС». При этом организуется дополнительный диалог для получения разного рода характеристик. В случае отрицательного ответа происходит переход к следующей начальной вершине в КлСемС, не совпадающей с текущей и так пока не закончится информация в КлСемС. Если всегда был отрицательный ответ, то надо обратиться к казуальной семантической сети, пример которой привден ниже:

(1(Куб())(Вершина())()) ;Если есть куб, то обязательно должна быть вершина,т.е. куб является причиной существования вершины
(2(Электрическое поле())(Напряженность())()) ;Если есть электрическое поле, то обязательно должна существовать такая его характеристика, как напряженность
(3(Куб())(Центр())())
(4(Грань куба())(Центр())())

Пустые списки зарезервированы для внесения дополнительной информации. Последний пустой список зарезервирован для указания ссылок. Для приведенной казуальной семантической сети могут быть сформированы следующие ссылки:

(1(Куб())(Вершина())((3 4)())) ;Если есть куб, то обязательно должна быть вершина,т.е. куб является причиной существования вершины
(2(Электрическое поле())(Напряженность())()) ;Если есть электрическое поле, то обязательно должна существовать такая его характеристика, как напряженность
(3(Куб())(Центр())((1 4)(4)))
(4(Грань куба())(Центр())(()(3)))

Ссылка как список имеет такую структуру: ((Список номеров списков, где встречается начальная вершина)(Список номеров списков, где встречается конечная вершина)). В данном случае в казуальной семантической сети отсутствует явнвя подсказывающая информация. С помощью казуальной семантической сети образовывать новые понятия. Для этого берется словосочетание конечной вершины и добавляется словосочетание начальной вершины в родительном падеже. Например: вершина куба, напряженность электрического поля и т.д. Это возможно не всегда. Семантическая сеть шаблонов задач представляет собой следующую структуру:

(1(Описание шаблона 1-й задачи)((Список списков ссылок на подобные задачи)(Ссылка на решение задачи)))
(2(Описание шаблона 2-й задачи)((Список списков ссылок на подобные задачи)(Ссылка на решение задачи)))
…..
(n(Описание шаблона n-й задачи)((Список списков ссылок на подобные задачи)(Ссылка на решение задачи)))

Внутри текста может быть ссылка на задачи от понятия в виде /относительный номер начального слова понятия – относительный номер последнего слова понятия / @ [список

номеров списков шаблонов задач]. Ссылка всегда находится после последнего слова понятия, таких ссылок может быть несколько. Подобность задачи отражает некоторую идентичность(схожесть) по структуре, по классу задач и т.д. Построение шаблона задачи напоминает построение схемы программы.

Для примера рассмотрим шаблон приведенной задачи.

(В вершинах X_5, X_2, X_4 куба $X_1X_2X_3X_4X_5X_6X_7X_8$, ребро которого равно a, помещены точечные заряды q. Выразить результирующую ?ХАРАКТЕРИТИКУ? создаваемого ими ?НЕКОТОРОГО? поля в точках X_1 и X_7 через вектор X_1X_7. Найти абсолютную величину ?ХАРАКТЕРИСТИКИ? в точках X_3 и X_6, в центре грани $X_5X_6X_7X_8$ и в центре куба. ?ХАРАКТЕРИТИКА? определяется по следующему правилу. Если в точке О находится точечный заряд q, то ?ХАРАКТЕРИТИКА? **Y**, создаваемого им ?НЕКОТОРОГО? поля в точке М выражается формулой **Y** =f(k,OM,**OM)**, где k зависит от выбора системы единиц).

При составлении шаблона данной задачи произошло абстрагирование от конкретных обозначений вершин куба. В подсистеме приобретения знаний использование переменных аналогично исчислению предикатов первого порядка, т.е. переменные x, y. z, X, Y, Z как с индексами, так и без индексов обозначают в шаблонах задач произвольные конкретные имена переменных в случае прописных символов и предметных переменных в случае строчных символов. Кроме этого произошло абстрагирование от наименования конретной характеристики конкретного поля. Такая абстракция показывается с помощью некоторого ключевого слова, изображенного прописными буквами и ограниченного с обоих сторон знаками вопроса. С помощью функционального символа f произошло абстрагирование от конкретной формулы. Текст приведенного шаблона задачи не является полной абстракцией задачи. С помощью данного примера показан один из способов получения шаблона задачи.

Рассмотрим функцию анализа словосочетаний и шаблонов задач. В семантической сети задачи ЗдчСемс1 в одной из конечных вершин имеется словосочетание «Формулой **E** = (k*q)/OM^3***OM**», которое понятием назвать нельзя. Оно состоит из понятия «формула» и обозначения «**E** = (k*q)/OM^3***OM**». Возникает вопрос каков должен быть критерий различения понятий просто от словосочетаний. На первом уровне разработки подсистемы решения об отнесении словосочетания к понятию или нет можно возложить на квалифицированного пользователя. Ясно, что это не являетсяя решением возникшей проблемы. Скорее всего большинство словосочетаний, не являющихся понятиями, можно разбить на понятия и словосочетания программным путем, но останутся такие ситуации, когда решение должен принимать пользователь. В идеальном, но не реальном, случае должна быть полная информация в различного рода справочниках.

Для анализа шаблонов нужно уметь исходную задачу преобразовывать в шаблон и потом искать подходящий. Можно искать похожесть по конкретным сочетаниям и, найдя некоторый шаблон, попытаться с помощью такой подсказки преобразовать исходную задачу.

В данном докладе не отражены действия по нахождению источников знания, по созданию шаблонов и по организации записи полученной информации в нужное место. Это связано с наличием нескольких вариантов решения указанных проблем, и с трудностями выбора конкретного решения.