

Обробка інформації в складних організаційних системах

УДК 519.7

Н.А. Валенда

Харьковский национальный университет радиоэлектроники, Харьков

МЕТОД ФОРМАЛИЗАЦИИ СЕМАНТИЧЕСКОГО СЛОВАРЯ НА ОСНОВЕ ФУНКЦИЙ

В работе предложена модель формализации семантики естественного языка в виде суперпозиций семантических функций. Рассмотрен состав и структура семантических функций. Данная модель позволяет учитывать многозначность языковых единиц. Рассмотрена модель формализации семантического словаря на основе функций.

Ключевые слова: словарь, семантический анализ, контекст, семантическая функция, лингвистический процессор.

Введение

Степень формализации языка определяется степенью формализации его семантики. По отношению к естественным языкам это утверждение приводит к важным последствиям. Так как в основе семантической структуры ЕЯ лежит модель окружающего нас мира, то возможность точного описания ЕЯ прямо зависит от возможности построения точных моделей реальной действительности. Ясно, что реальная действительность неформализуема, следовательно, казалось бы, абсурдна даже постановка вопроса о возможности формализации ЕЯ. Но это неправильный вывод.

Для формализации семантики ЕЯ достаточно языковой модели мира, а эта модель содержится в словарном составе ЕЯ. Можно говорить о формальной модели естественного языка, не выходя за его рамки [1].

В данной работе рассматривается функциональный подход к формализации языка. Связи между словами можно представить ограниченным набором функций. Существует набор семантических отношений, которые используются в большинстве работ для описания отношений между элементами предложений. Этот набор несколько варьируется в различных работах, но основной перечень отношений является неизменным. Если задать отношения, существующие в предложении, в виде функций, то получим суперпозицию функций, поскольку функции зависимых элементов предложения будут встраиваться в функции главных как аргументы. Слова тоже рассматриваются как функции, которые принимают одно из значений, соответствующих слову в языке, в зависимости от контекста.

В работе рассматривается аппарат семантических функций и формализованное представление семантического словаря.

Семантический словарь

Основой для формальной модели семантики выступает толковый словарь, который содержит лексические значения слов. Слово становится основной информационной единицей, а словарь – основной работы лингвистической системы. Любой словарь может быть представлен как лексикографическая система [2, 3].

Обозначим как $W(L)$ конечное множество всех слов, входящих в язык L . Пусть слово x , $x \in W(L)$ имеет в языке L несколько значений. Для установления связи между словом x и его значением введем функцию значения слова – $V(x)$. Областью определения для функции $V(x)$ является $W(L)$, областью значений – концепты предметной области.

$$V(x) = \{\alpha_1, \dots, \alpha_m\}, \quad (1)$$

где $\alpha_1, \dots, \alpha_m$ – значения, приписываемые слову x .

Каждому значению слова соответствует статья толкового словаря. Формально это можно записать следующим образом:

$$V_1(x) = \alpha_1,$$

...

$$V_i(x) = \alpha_i,$$

...

$$V_m(x) = \alpha_m.$$

Для отображения $x \rightarrow \alpha_i$, $i = \overline{1, m}$ необходимы знания, позволяющие сделать выбор. Вне кон-

текста такое значение определить невозможно, поэтому каждому значению слова сопоставляется описание некоторой ситуации, при которой слово принимает данное значение.

Для описания ситуации необходимо указать множество объектов, принимающих в ней участие, их связи и взаимное влияние. Построение соответствующих формализмов давно является предметом изучения компьютерной лингвистики. Система падежей, введенная Филмором и дополненная его последователями, является одним из наиболее распространенных способов описания семантики языковых конструкций [4].

Для описания событий удобно использовать функциональную структуру естественного языка. В качестве функций рассматриваются отношения, которые существуют между объектами. Будем называть такие функции семантическими, поскольку они позволяют описать ситуацию, используя семантические категории объекта, действия, их признаков.

Семантические функции

Семантическая функция имеет вид:

$$f_i(x_1^1, \dots, x_n^1), \quad (2)$$

где i – задает номер семантического отношения, $i = \overline{1, 16}$;

x_j^k – аргумент функции;

k – индекс, указывающий роль данного аргумента в функции, $k = \overline{1, 10}$.

Для каждого аргумента функции указывается роль, в которой он выступает в функции, в аргументе x_j^k роль задает индекс k . В данной работе взят за основу перечень ролей или семантических валентностей, выделенных в работах Филмора для глагольных функций, который был дополнен новыми ролями. Это связано с расширением перечня функций.

Список ролей, используемых для аргументов функций: k_1 = субъект; k_2 = действие; k_3 = контрагент; k_4 = объект; k_5 = адресат; k_6 = пациент; k_7 = результат; k_8 = инструмент; k_9 = признак; k_{10} = значение.

В данной работе используются следующие семантические функции:

f_1 = действие; f_2 = способ; f_3 = принадлежность; f_4 = признак; f_5 = состояние; f_6 = средство; f_7 = количество; f_8 = назначение; f_9 = направление; f_{10} = степень; f_{11} = место; f_{12} = содержание; f_{13} = причина; f_{14} = время; f_{15} = период; f_{16} = часть.

Ниже приведен вид соответствующих функций:

$$f_1(x_1^{k_2}, x_2^{k_1}, x_3^{k_4}, x_4^{k_3}, x_5^{k_5}, x_6^{k_6}, x_7^{k_7}, x_8^{k_8});$$

$$f_2(x_1^{k_2}, x_2^{v_1}), \quad v_1 = k_4 \vee k_2;$$

$$f_3(x_1^{k_4}, x_2^{k_4});$$

$$f_4(x_1^{v_1}, x_2^{k_9}, x_3^{k_{10}}), \quad v_1 = k_4 \vee k_1 \vee k_2;$$

$$f_5(x_1^{v_1}, x_2^{k_{10}}); \quad v_1 = k_2 \vee k_4;$$

$$f_6(x_1^{k_2}, x_2^{k_4});$$

$$f_7(x_1^{k_4}, x_2^{k_{10}});$$

$$f_8(x_1^{v_1}, x_2^{v_2}), \quad v_1 = k_2 \vee k_4, \quad v_2 = k_2 \vee k_4;$$

$$f_9(x_1^{v_1}, x_2^{k_4}), \quad v_1 = k_2 \vee k_4;$$

$$f_{10}(x_1^{v_1}, x_2^{k_{10}}), \quad v_1 = k_2 \vee k_4;$$

$$f_{11}(x_1^{v_1}, x_2^{k_4}), \quad v_1 = k_2 \vee k_4;$$

$$f_{12}(x_1^{k_4}, x_2^{v_1}), \quad v_1 = k_2 \vee k_4;$$

$$f_{13}(x_1^{k_2}, x_2^{k_4});$$

$$f_{14}(x_1^{v_1}, x_2^{k_{10}}), \quad v_1 = k_2 \vee k_4;$$

$$f_{15}(x_1^{v_1}, x_2^{k_{10}}), \quad v_1 = k_2 \vee k_4;$$

$$f_{16}(x_1^{k_4}, x_2^{k_4}, \dots, x_n^{k_4}).$$

Перечень функций может быть расширен при необходимости.

В семантической функции аргументы x_j^k могут задаваться или функциями значения слов, или семантическими функциями. Возможные значения аргумента x_j^k :

$$x_j^k = \begin{cases} V_i^k(x_j), \\ f_m^k(x_1^1, \dots, x_n^k). \end{cases} \quad (3)$$

Аргументом функции может являться переменная, не имеющая значения. Переменные обозначаются большими латинскими буквами. Значения переменных определяются в соответствии с (3). Множество возможных значений переменной может быть ограничено. Для этого вводятся множества признаков, которым должно отвечать возможное значение:

$M(P_{\text{morph}})$ – множество морфологических признаков;

$M(P_{\text{sem}})$ – множество семантических признаков.

$$M(P_{\text{morph}}) = \{p_1, \dots, p_{14}\}, \quad (4)$$

где p_1 =часть речи, p_2 =одушевленность, p_3 =род, p_4 =число, p_5 =падеж, p_6 =степень сравнения, p_7 =краткость, p_8 =репрезентация, p_9 =вид, p_{10} =время, p_{11} =лицо, p_{12} =пассивность, p_{13} =разряд местоимений, p_{14} =тип числительных.

$$M(P_{sem}) = \bigcup_{i=1}^N \alpha_i, \quad (5)$$

где α_i – концепты.

Переменная имеет вид: $X_{[P_{morf}, P_{sem}]}$, где

$$P_{morf} \subset M(P_{morf}) \vee P_{morf} = \emptyset, \\ P_{sem} \subset M(P_{sem}) \vee P_{sem} = \emptyset.$$

Введение переменных позволяет использовать семантические функции как некоторую рамку для заполнения элементами, удовлетворяющими заданным ограничениям.

Дадим определение семантической функции.

Термом является:

- переменная;
- слово, заданное функцией значения слова,

если x_j – слово, то $V_1(x_j)$ – функция значения слова;

- семантическая форма.

Семантическая форма – это функциональная константа, соединенная с подходящим числом термов. Если f_j – функциональная m -местная константа,

x_j – термы, $j = \overline{1, m}$, то соответствующая форма имеет вид $f_j(x_1, \dots, x_m)$.

Семантическая функция – это функциональная константа, соединенная с подходящим числом термов. Если f_i – функциональная n -местная константа,

x_i – термы, $i = \overline{1, n}$ то семантическая функция имеет вид $f_i(x_1, \dots, x_n)$.

Суперпозиция семантических функций задает ситуацию, в которой слово x примет значение α_i , поэтому можно провести соответствие между значением слова и суперпозицией, соответствующей этому значению. Получим отношение зависимости значения слова от суперпозиции семантических функций.

$$V_i(x) = \alpha_i, \\ V_i(x) \rightarrow f_j(y_1, \dots, y_n).$$

Как было сказано выше, α_i , $i = \overline{1, N}$ является концептом предметной области. Сами по себе концепты как изолированные объекты не расширяют объем знаний о предметной области. Важной частью любого способа представления знаний является классификация объектов по категориям. Хотя взаимодействие с миром происходит на уровне отдельных объектов, формирование рассуждений происходит в основном на уровне категорий [5].

Категории служат для организации и упрощения базы знаний с помощью наследования. Подклассы наследуют свойства классов. Отношения

между классами и подклассами позволяют организовывать категории в виде некоторой таксономии. Формально ввести понятие таксономии можно на основании более общего понятия онтологии [6].

Формальная модель онтологии – упорядоченная тройка вида:

$$O = \langle X, \mathfrak{R}, \Phi \rangle, \quad (6)$$

где X – конечное множество концептов (понятий, терминов) предметной области, которую представляет онтология O ;

\mathfrak{R} – конечное множество отношений между концептами (понятиями, терминами) предметной области;

Φ – конечное множество функций интерпретации, заданных на концептах и/или отношениях онтологии O .

Введя ограничения в общее определение онтологии можно получить таксономию:

$$O = T^0 = \langle X, \mathfrak{R}, \{ \} \rangle, \\ X = \bigcup_{i=1}^N \alpha_i, \\ \mathfrak{R} = \{ is_a \}.$$

Под таксономической структурой понимается иерархическая система понятий, связанных между собой отношением is_a – быть элементом класса. Отношение is_a имеет фиксированную заранее семантику и позволяет организовать структуру понятий в виде дерева.

К системам, предназначенным для организации и проведения рассуждений, касающихся категорий, относятся семантические сети, которые предоставляют графические средства, позволяющие визуальное представить базу знаний, и эффективные алгоритмы для логического вывода сведений о любом объекте на основании его принадлежности к некоторой категории [5].

Семантическая сеть состоит из множества концептуальных графов, представляющих логические формулы. Она позволяет визуализировать множество отношений между концептуальными графами, ее составляющими.

Концептуальный граф – это конечный, связанный, двудольный граф. Узлы графа определяют понятия, или концептуальные отношения. В концептуальных графах метки дуг не используются. Отношения между понятиями представляются узлами концептуальных отношений. Каждый концептуальный граф представляет одно высказывание. Графы могут быть произвольной сложности, но они должны быть конечными.

Для концептуальных графов вводится иерархия типов, которая позволяет осуществлять множественное наследование. Иерархия типов – это частичное упорядочение на множестве типов, которое

можно обозначить символом \leq . Если s и t – есть типы и $t \leq s$, то говорят, что t – это подтип, а s – супертип. Поскольку эта систематизация частична, тип может иметь один или несколько супертипов, а также один или несколько подтипов.

Можно сделать вывод, что семантическая сеть – это таксономия, узлами которой являются концептуальные графы. При необходимости можно расширить возможности сети за счет расширения множества \mathcal{R} .

Узлами семантической сети являются α_i , каждому из которых соответствует суперпозиция семантических функций, задающая ситуацию, в которой слово x принимает i -е значение. Следовательно, семантический словарь может рассматриваться как семантическая сеть. Если знаний, хранящихся в такой сети, оказывается не достаточно для анализа текстов ЕЯ, то возможно расширение знаний о предметной области за счет сопоставления концептуальных графов узлам сети.

Семантический словарь S можно задать в виде объединения множеств $V_i(x)$ для всех $x \in W(L)$:

$$S = \bigcup_{j=1}^N \bigcup_{i=1}^m V_i(x_j). \quad (7)$$

Записи семантического словаря S_k будут иметь следующий вид.

$$S_k = V_i(x) = x \cup f_m(x_1^1, \dots, x_n^1) \cup \alpha \cup k \cup M(u), \quad (8)$$

где x – заглавное слово;

$f_m(x_1^1, \dots, x_n^1)$ – суперпозиция семантических функций, соответствующая значениям слова x ;

α – ссылка на иерархию типов;

k – ссылка на концептуальный граф или пустое множество;

$M(u)$ – множество ссылок на словарь устойчивых словосочетаний или пустое множество.

Выводы

Результаты данной работы могут быть использованы для построения формального представления в системах автоматического перевода, для систем с естественно-языковым интерфейсом, в поисковых системах. Основным преимуществом данного подхода является детальная обработка многозначных слов, что позволяет существенно сократить смысловой недетерминизм для большинства слов и словосочетаний в результирующем формальном представлении.

Список литературы

1. Тузов В.А. Компьютерная лингвистика. Опыт построения компьютерных словарей [Текст] / В.А. Тузов. – СПб.: Изд. СПбГУ, 2002. – 650 с.
2. Широков В.А. Элементы лексикографии [Текст] / В.А. Широков. – К.: Довіра, 2005. – 304 с.
3. Широков В.А. Феноменологія лексикографічних систем [Текст] / В.А. Широков. – К.: Наукова думка, 2004. – 327 с.
4. Филмор Ч. Дело о надежде [Текст] / Ч. Филмор // Новое в зарубежной лингвистике. Вып. X. Лингвистическая статистика. – 1981. – С. 369-495.
5. Рассел С. Искусственный интеллект: современный подход [Текст] / С. Рассел, П. Норвиг. – М.: Вильямс, 2006. – 1408 с.
6. Гаврилова Т.А. Базы знаний интеллектуальных систем [Текст] / Т.А. Гаврилова, В.Ф. Хорошевский. – СПб.: Питер, 2000. – 384 с.

Поступила в редколлегию 9.07.2013

Рецензент: д-р физ.-мат. наук, проф. А.В. Грицунов, Харьковский национальный экономический университет, Харьков.

МЕТОД ФОРМАЛІЗАЦІЇ СЕМАНТИЧНИХ СЛОВНИКА НА ОСНОВІ ФУНКЦІЙ

Н.А. Валенда

У статті запропоновано модель формалізації семантики природної мови у вигляді суперпозицій семантичних функцій. Дана модель дозволяє враховувати багатозначність мовних одиниць.

Ключові слова: словник, семантичний аналіз, контекст, семантична функція, лінгвістичний процесор.

METHOD FORMALIZATION SEMANTIC DICTIONARY BASED ON THE FUNCTION

N.A. Valenda

The model of formalization of semantics of natural language designs as superpositions of semantic functions which allows quality improvement of systems of automatic processing of texts at multivalued language units analysis is developed in the thesis.

Keywords: dictionary, semantic interpretation, context, semantic function, linguistic processor.