УДК 519.62

НЕКОТОРЫЕ АСПЕКТЫ ФОРМАЛИЗАЦИИ ТОЛКОВОГО СЛОВАРЯ С ПРИМЕНЕНИЕМ Λ-ИСЧИСЛЕНИЯ

Г.Ф. Дюбко
Кандидат технических наук, профессор*
Контактный тел.: 702-14-46
E-mail: prof_dubko@yandex.ru

М.Ю. Дзёх*
Контактный тел.: 050-767-48-03
E-mail: mary.dzyoh@gmail.com

Д.В. Преснякова
Аспирант*
*Кафедра «Программное обеспечение электронных вычислительных машин»
Харьковский национальный университет радиоэлектроники
ул. Серповая, 2, г. Харьков, 61166
Контактный тел.: 702-14-46
E-mail: Darya_Presnyakov@mail.ru

Побудовано модель опису контексту за допомогою лямбда-числення і запропоновано механізм обчислення словникової статті, який базується на семантичних функціях, семантичних примітивах і результатах синтаксико-семантичного аналізу речення

Ключові слова: семантична функція, семантичний примітив, λ-числення, λ-формула

Построена модель описания контекста с помощью лямбда-исчисления и предложен механизм вычисления словарной статьи, который базируется на семантических функциях, семантических примитивах и результатах синтаксико-семантического анализа предложения

Ключевые слова: семантическая функция, семантический примитив, λ-исчисление, λ-формула

This article represents the model of context description with the use of lambda calculus and the mechanism of vocabulary entry calculation, which is based on semantic functions, semantic primitives and the results of analysis of an examined sentence

Key words: semantic function, semantic primitive, lambda calculus, lambda formula

1. Введение

Проблема интеллектуального анализа и понимания текстов на естественном языке появилась одновременно с созданием компьютеров. Однако до настоящего времени достичь полного успеха в её решении не удалось.

Методы для решения частных задач разработаны в большом количестве, но количество так и не перешло в качество[1].

Накопление огромного количества информации в текстовом виде делает актуальным создание программ, ориентированных на ее обработку. Диапазон таких программ весьма широк: системы автоматического перевода, естественно-языковые интерфейсы, системы автоматического реферирования. Несмотря на широту диапазона решаемых задач, качество работы всех этих систем прямо зависит от уровня формализации естественного языка, который может предложить современная теория искусственного интеллекта. Большое количество современных работ в области искусственного интеллекта направлены на разработку моделей семантики, которые позволят сделать качественный скачок в семантической интерпретации текстов и улучшить результаты практической работы систем обработки текстовой информации [2].

Естественный язык (ЕЯ) обладает многими свойствами, необходимыми, для представления знаний. Но для того чтобы использовать его как средство представления знаний в ЭВМ, необходимо решить сложнейшую проблему содержательного (семантического) анализа текстов на ЕЯ. Вопросам формализации отдельных сторон ЕЯ посвящено немало работ как у нас в стране, так и за рубежом. Большинство работ по формализации семантики языка состоит в разработке формальных моделей и языков их представления, которые должны позволять одинаковое формальное каноническое представление для различных естественно-языковых конструкций (ЕЯК), имеющих одинаковый смысл.

2. Постановка задачи

В языковой практике человеческой коммуникации существенное место занимают толковые словари, где каждому слову сопоставляются его толкования в зависимости от окружающего контекста этого слова. Эти толкования можно трактовать как значение слова. Каждое толкование называют словарной статьей. В общем случае каждому слову соответствует более одной словарной статьи, что приводит к неоднозначности толкования.

Информация, содержащаяся в толковых словарях, предназначена для обработки человеком и представлена на том же самом языке, который она описывает. Смысловые аспекты, описываемые в толковых словарях, называются лексической семантикой. Лексическая семантика играет существенную роль в понимании смысла ЕЯК, и поэтому естественно предполагать, что лексическая семантика должна быть использована в системах автоматического анализа языка.

Использование лексической семантики при автоматическом анализе ЕЯК требует формализации толкового словаря, а это, в свою очередь, предполагает наличие формального языка описания словарных статей и механизмов вычисления значений слов в зависимости от окружающего контекста.

В качестве формального языка описания словарных статей в рассматриваемой работе выбран язык семантических функций, а в качестве механизма вычисления соответствующей словарной статьи – лямбда-исчисление (λ-исчисление). Формальный толковый словарь на основе семантических функций и λ-исчисления назовем семантическим словарем.

Под естественно-языковыми конструкциями, рассматриваемыми в данной работе, понимаются словосочетания ЕЯ. Простые повествовательные предложения являются особой формой словосочетания. Слова, составляющие ЕЯК, получают свои значения в рамках предложения. Рассматриваются предложения, главным словом которых является глагол, обозначающий действие (как семантический примитив), а остальные словосочетания предложения образуют периферию.

Целью данной работы является разработка принципов создания семантического словаря (λ-словаря) и механизма вычисления словарной статьи для глаголов, обозначающих действие, на основе λ-исчисления.

3. Семантические функции

Элементарным понятием естественного языка является словосочетание с законченным смыслом. Это может быть простое или сложное предложение, словосочетание из двух слов, текст. Словосочетание - это синтаксическая конструкция, образующаяся на основе подчинительных связей: согласования, управления и примыкания. В словосочетание полностью переносятся все те отношения, те общие и частные, более конкретные значения, которые возникают при подчинительной связи: эти отношения представляют собой значения словосочетаний. В составе словосочетания выделяются компонент стержневой (или главенствующий) и компонент зависимый (компоненты зависимые): стержневым компонентом является грамматически главенствующее слово, своими лексико-грамматическими свойствами предопределяющее связь; зависимым компонентом - форма слова (формы слов), грамматически подчиненная (подчиненные)[3]. В данной работе мы ограничимся описанием простых повествовательных предложений.

Чтобы сделать возможной компьютерную обработку естественно-языковых конструкций (словосочетаний), необходимо формализовать представление их смысла. Такая формализация возможна с применением семантических функций [4]. Посредством семантических функций может быть описан поверхностный смысл конструкций ЕЯ.

Дадим определение семантической функции. Пусть V – семантическая функция слова ЕЯ, F – семантическая функция словосочетания. Функция V имеет вид $V_i(\alpha)$, где α – слово естественного языка в канонической форме, i – № словарной статьи из семантического словаря.

Например, в [5] слову «абажур» в первой словарной статье соответствует смысл «колпак для лампы», поэтому будет справедливым равенство V_1(«абажур»)= «колпак для лампы».

Словосочетание ЕЯ, состоящее из слов (α_1, α_2,..., α_n), представляет семантическую функцию F_имя(X_1, X_2, ..., X_n), где $X_j = V_i(\alpha_k)$, или X_j = F'_имя, где k – номер словарной статьи в семантическом словаре; имя – название функции, которое соответствует названию словосочетания; F'_имя – семантическая функция, вложенная в F_имя, α_k – слово из словосочетания.

Например, словосочтанию «колпак для лампы» соответствует семантическая функция F_предлог(V_i(«для»),V_j(«колпак»),V_k(«лампы»)), где i, j и k – обозначают номер словарной статьи в семантическом словаре. Значениями этих функций является смысл, который можно сопоставить слову или словосочетанию.

Приведем примеры семантических функций для словосочетаний в русском языке. Пусть рассматривается словосочетание «портфель профессора». Ему соответствует семантическая функция F_генитив(V_i(«портфель»), V_j(«профессора»)). Для словосочетания «количество комнат в доме» семантическая функция имеет вид: F_генитив(V_i(«количество»), F_предлог(V_j(«в»), V_k(«комнат»), V_r(«доме»))).

Областью определения семантической функции являестя множество слов, входящих в словосочетание. Аргументами семантической функции могут быть другие семантические функции, имеющие свою область определения. Задавая правила вычисления семантических функций, можно получать их значения.

4. Семантические примитивы

Знания, содержащиеся в семантическом словаре, выражены посредством семантических единиц (элементарных смыслов), с помощью которых представлена и сама семантическая структура ЕЯ-конструкций. Простейшей семантической единицей служит семантический примитив, выражающий значение единицы в реальном мире. Лексическая семантика базируется на множестве таких примитивов – элементарных единиц смысла. Элементарные смыслы могут обра-

зовать иерархическую структуру, на верхнем уровне которой находятся примитивы ОБЪЕКТ, ДЕЙСТВИЕ, АБСТРАКЦИЯ, ОТНОШЕНИЕ. Каждому из этих примитивов соответствуют примитивы более низкого уровня [6]. Например, класс семантических примитивов ОТНОШЕНИЕ можета содержать такие семантические единицы, как ПРИЧИНА, СЛЕДСТВИЕ, РОДСТВО, ПРИНАДЛЕЖНОСТЬ, и т.п.

А.Вежбицкая [7], например, представляет множество семантических примитивов в виде неструктурированого множества или сетки категорий. Ею было видвинуто около шестидесяти кандидатов в универсальные смыслы, например: Субстативы:{Я, ТЫ, НЕКТО/ЛИЦО, НЕЧТО/ВЕЩЬ, ЛЮДИ, ТЕЛО}, Детерминаторы:{ЭТОТ, ТОТ ЖЕ, ДРУГОЙ}, Кванторы:{ОДИН, ДВА, НЕСКОЛЬКО/НЕМНОГО, ВЕСЬ/ВСЕ, МНОГО/МНОГИЕ}, Атрибуты: {ХОРОШИЙ, ПЛОХОЙ, БОЛЬШОЙ, МАЛЕНЬКИЙ} и т.д. Семантические примитивы только способствуют описанию семантической структуры текста. Отталкиваясь от семантических примитивов верхнего уровня можно построить иерархию для определенной предметной области, при надобности расширяя и дополняя ее.

Семантически словарная статья будет описана следующим образом. Каждому слову α языка сопоставляется семантическая функция V_i (α), где i – номер словарной статьи слова α в семантическом словаре. Значением функции V_i является имя семантического примитива, говорящее о смысле слова α в употребляемом контексте. Семантика словосочетаний выражается семантической функцией F_имя, состоящей из более простых семантических функций: F_имя(X_1,...,X_n), где F_имя - наименование семантической функции, $X_i(1\le i\le n)$ – семантическая функция V_i (α).

Используя семантические функции и примитивы, соответствующие аргументам этих семантических функций можно вычислять примитивы, являющиеся семантическим значением семантической функции.

5. Λ-словарь

Лямбда-исчисление, предложенное А. Черчем, является теоретической основой описания вычислительных процессов, которая не содержит в явном виде понятие ячеек памяти для хранения значений переменных и последовательности вычислений как процесса изменения состояния памяти. А. Черч построил систему, где используются правила преобразований, с помощью которых можно получать из одних функций другие, эквивалентные им. Λ-исчисления рассмотрены в [8]. Λ-исчисление позволяет объединение функционального и предикатного подходов, т.е. в λ-формулы можно встраивать формулы исчисления предикатов.

Объединение λ-формулы с предикатными формулами и семантическими примитивами позволяет создать механизм для вычисления словарной статьи глагола и его периферии (употребление глагола в конкретном предложении) исходя из результатов синтаксико-семантического анализа предложения. Для этого с каждой словарной статьей глагола α связывается соответствующим образом построенная λ-формула. В общем случае такая λ-формула имеет вид:

$$\lambda x_1.\lambda x_2.\lambda x_n.(P_1(X_1)\&P_2(X_2)\&...\&P_m(X_m)\&Q_1(X_1,...,X_n)\&...\&Q_k(X_1,...,X_n))@\omega_1...@\omega_n \quad (1)$$

где: $P_i(X_i)$, $(1\le i\le n)$, $Q_i(X_1,...,X_n)$ $(1\le i\le m)$ – предикаты;

имя предиката P_i – семантический примитив;

предикаты Q_i – формализуются как предикаты равенства морфологических признаков соответствующих аргументов каким-то вполне определенным значениям;

$\omega_1,...,\omega_n$ – периферия, которая определяется результатом синтаксико-семантического анализа глагола α в конкретном предложении, т.е. $\omega_1,...,\omega_n$ – это семантические функции для внутренних словосочетаний предложения.

Предикаты P_i и Q_j принимают значения «истина», если семантический примитив x_i (аргумент P_i) равен P_i, а для Q_j выполняются соответствующие равенства. В целом истинность λ-формулы определяется конъюнкцией ее составляющих (предикатов) после выполнения редукции, т.е. подстановки $\omega_1,...,\omega_n$ вместо аргументов в (1).

С каждой словарной статьей глагола α связана одна или более формул типа (1). Выбирается та словарная статья, для которой формула типа (1) становится истинной после редукции λ-формулы.

Отметим, что множество $\{\omega_1,\omega_2,...,\omega_n\}$ в (1) рассматривается как упорядоченное множество, т.е. предполагается фиксированный порядок слов в анализируемом предложении, чего нет в реальной языковой практике. Поэтому в процессе вычисления соответствующей словарной статьи необходимо рассматривать все перестановки множества $\{\omega_1,\omega_2,...,\omega_n\}$ (с изменением порядка следования словосочетаний), и выбирается та перестановка, которая дает истинный результат. С каждой словарной статьей связан вполне определенный вид семантической функции с аргументами, обозначенными переменными из (1), в которую вместо переменных подставляются значения $<\omega_1, ..., \omega_n>$. Эта семантическая функция выбирается в качестве значения анализируемого предложения.

Формирование λ-формулы типа (1) можно осуществить путем анализа и формализации словарных статей глагола α в толковом словаре.

Рассмотрим на конкретных примерах формирование формул для глаголов, обозначающих действия.

Для глагола «работать» в значении «находиться в действии» λ-формула будет иметь вид:

$\lambda x.$(ОБЪЕКТ(x))@ω.

Это справедливо для предложений «Машина работает», «Завод работает».

Для «работать» в значении «осуществлять деятельность» λ-формула быдет выглядеть следующим образом:

$\lambda x_1.\lambda x_2.$(ОБЪЕКТ(x_1)&

ОБЪЕКТ(F_предлог(V(«над»),

V(x_2)))&x_1.одушевленность=«одушевленный»)@ω_1@ω_2.

Это справедливо для предложений типа «Работать над рукописями».

Для «работать» в значении «иметь постоянное занятие» λ-формула будет выглядеть следующим образом:

$\lambda x_1.\lambda x_2.$(ОБЪЕКТ(x_1)&СПЕЦИАЛЬНОСТЬ(x_2)&x_1.одушевленность=«одушевленный»&x_2.падеж=«предложный»).

Это справедливо для предложений типа «Работать слесарем», «Работать дворником».

Для глагола «идти» в значении «двигаться, переступая ногами» λ-формула будет выглядеть следующим образом:

$\lambda x_1.\lambda x_2.$(ОБЪЕКТ(x_1)&МЕРА(x_2)&x_1.одушевленность=«одушевленный»&x_2.падеж =«творительный»)@ω_1@ω_2.

Это справедливо для предложений типа «Лошадь идет шагом».

Сформированная вышеописанным способом информация содержится в семантическом словаре, названным λ-словарем. Здесь каждой словарной статье конкретного глагола приписана λ-формула типа (1). Эта формула становится истинной, если периферия глагола совпадает с периферией в λ-формуле. Информацию, представленную в λ-словаре можно структурировать так, как показано на рис. 1.

Каждой словарной статье глагола может соответствовать одна или несколько λ-формул, а также две семантические функции. Одна из семантических функций представляет собой фрейм для построения смыслового эквивалента анализируемого предложения (с точки зрения лексической семантики). Другая семантическая функция представляет значение глагола, т.е. его толкование в формальном языке.

Кроме λ-словаря при анализе исходных предложений нужна информация о семантических примитивах, глаголе и его периферии, что требует механизмов их вычислений. Одним из возможных подходов к вычислению семантических примитивов является организация примитивов в онтологию и использование значений примитивов в семантических функциях. Вычисление периферии глагола можно осуществить, базируясь на методах синтаксического анализа.

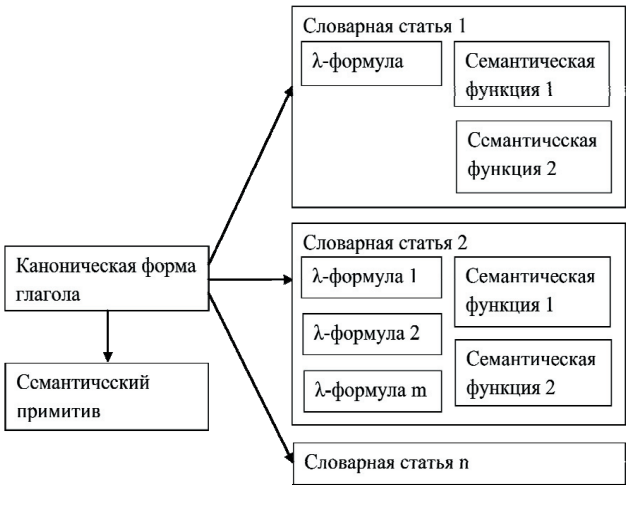

Рис. 1. Структурирование информации в λ-словаре

Целью анализа предложения на основе λ-словаря является построение семантической функции предложения, соответствующей вычисленной словарной статье.

Схема вычисления словарной статьи и соответствующей ей семантической функции представлена на рис. 2. Здесь α – анализируемый глагол, $\{\omega_1,..., \omega_n\}$ – его периферия, $<\beta_1, ..., \beta_n>$ – упорядоченная n-ка из множества $\{\omega_1,..., \omega_n\}$, выбираемая для редукции λ-формулы. FORM – формула логики предикатов, полученная в результате редукции.

Заметим, что семантическая функция являющаяся результатом вычисления, отражает лексическую семантику анализируемого предложения. Кроме нее определяется еще одна семантическая функция (см. рис. 1), которая является лексическим значением употребляемого глагола.

Последняя семантическая функция может трактоваться как обобщенный смысл анализируемого предложения.

Рассмотрим определение соответствующей словарной статьи для глагола «идти» на примере простого предложения «Мальчик идет в школу». Результатом синтаксического разбора будет ω_1=«мальчик», ω_2=«F_предлог(V(«в»), V(«школа»))» (слово «школу» приводится к канонической форме с помощью морфологического словаря).

Рис. 2. Схема вычисления словарной статьи и ее семантической функции

Глаголу «идти» в значении «направляться куда-либо» соответствует λ-формула:

$\lambda x_1.\lambda x_2$(ОБЪЕКТ(x1)&ПЕРЕМЕЩЕНИЕ(F_предлог(V(«в»), ОБЪЕКТ(x_2)))& x_1.одушевленность=«одушевленный»&x_2.падеж=«предложный»)@ω_1@ω_2

Подставим ω_1 и ω_2 в λ-формулу и получим:

$\lambda x_1.\lambda x_2$(ОБЪЕКТ(«мальчик»)&ПЕРЕМЕЩЕНИЕ(F_предлог(V(«в»), ОБЪЕКТ(«школа»)))&мальчик.одушевленность=«одушевленный»&школа.падеж=«предложный»).

Сравнение семантических примитивов и морфологических признаков слов дает true.

Следует заметить, что семантическая функция F_предлог будет давать разные семантические примитивы, в зависимости от того, с каким зависимым словом употребляется предлог.

Если рассматривается семантическая функция F_предлог, нужно уметь декомпозировать эту функцию на составляющие для обеспечения редукций.

Например, для F_предлог(V(«в»),V («школа»)) декомпозиция дает три компоненты: F_предлог, V(«в»), V(«школа»). Каждый декомпозированный элемент обозначается как ω_i, ω_{i+1}, ω_{i+2}.

Λ-формула для конструкций с предлогами должна выглядеть так (для нашего случая):

$\lambda x_1.\lambda x_2$(ОБЪЕКТ(X1)&F_предлог(F_предлог)&ПЕРЕМЕЩЕНИЕ(V(«в»))&ОБЪЕКТ(V(«X_2»))&X_1.одушевленность=одушевленный&X_2.падеж=«винительный»)@ω_1@ω_2@ω_3@ω_4

Можно предложить формулу и для других случаев. Например «Поезд идет в Киев». Если мы не различаем смыслы «идти шагом» и «двигаться», это будет одна общая λ-формула:

$\lambda x_1.\lambda x_2$(ОБЪЕКТ(X_1)&F_предлог(F_предлог)&ПЕРЕМЕЩЕНИЕ(V(«в»))&ОБЪЕКТ(V(«X_2»))&X_2.падеж=«винительный»)@ω_1@ω_2@ω_3@ω_4.

Если же мы хотим различить эти случаи, то вторая λ-формула будет отличаться от первой.

Выводы

Предложенный в работе подход позволяет создать формализированный толковый словарь. Этот словарь можно трактовать как базу лексических знаний, использование которых при автоматическом анализе естественно-языковых конструкций позволит повысить качество анализа за счет использования семантики.

В частности, λ-словарь используется в синтаксико-семантическом анализаторе, входом которого является предложения ЕЯ, а выходом семантическая функция.

Литература

1. Тузов, В.А. Языки представления знаний [Текст] / В.А. Тузов. – СПб. : СПбГУ, 1990. – 120 с.
2. Тузов, В.А. Компьютерная лингвистика. Опыт построения компьютерных словарей [Текст] / В.А. Тузов. – СПб. : СПбГУ, 2002. – 650 с.
3. Тестелец, Я.Г. Введение в общий синтаксис [Текст] / Я.Г. Тестелец. – М.: Изд-во РГГУ, 2001. – 800 с.
4. Дюбко, Г.Ф. Модель поверхностного смысла стественного языка на базе семантических функций [Текст] / Г.Ф. Дюбко, Д.В. Преснякова // Бионика интеллекта: научн.-техн. журнал. – 2007. – №1(66). – С. 103-106.
5. Ожегов, С.И. Словарь русского языка: Ок.57 тыс. слов [Текст] / Под ред.чл.-корр. АНСССР Н.Ю. Шведовой. –17-е изд., стереотип. – М.: Рус.яз.1985. – 797с.
6. Дюбко, Г.Ф. Формальная семантика и анализ естественного языка [Текст] / Г.Ф. Дюбко, Д.В. Преснякова // Бионика интеллекта: научн.-техн.журнал. – 2007. – №3(68). – С. 54-57.
7. Вежбицкая, Анна. Понимание культур через посредство ключевых слов [Текст] / Пер. с англ. А.Д.Шмелева. – М.: Языки славянской культуры, 2001. – 288 с.
8. Дюбко, Г.Ф. Формалізація семантики природної мови із використанням лямбда-числення [Текст] / Г.Ф. Дюбко, Д.В. Преснякова // Бионика интеллекта: научн.-техн. журнал. – 2007. – № 2(67). – С. 41-46.