



OSTIS-2011

(Open Semantic Technologies for Intelligent Systems)

УДК 004.823

АВТОМАТИЗАЦИЯ НАЧАЛЬНЫХ ЭТАПОВ ПРОЕКТИРОВАНИЯ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

А. В. Заболевая-Зотова, Ю. А. Орлова, А. С. Бобков (*yulia.orlova@gmail.com*)
*Волгоградский государственный технический университет,
г. Волгоград, Россия*

В данной работе описывается автоматизация начального этапа процесса проектирования. Предлагается методика анализа текста технического задания, которая состоит из трёх этапов: семантическая обработка текста, создание фреймовой структуры и создание диаграмм потоков данных системы, описанной в техническом задании.

Ключевые слова: автоматизация процесса проектирования, семантический анализ текста, фреймовые структуры.

Большинство работ в области САПР направлены на создание и совершенствование инструментария для автоматизации процесса проектирования. Наиболее известные из коммерческих программных продуктов, используемых для проектирования программного обеспечения, предназначены для визуализации промежуточных и конечных результатов процесса проектирования. Некоторые из них позволяют автоматизировать последние этапы проектирования: генерация программного кода, создание отчетной и сопровождающей документации и т.д. При этом задача автоматизации начального этапа проектирования – формирование и анализ текста технического задания – остается открытой.

Таким образом, актуальной является задача автоматизации начального этапа проектирования программного обеспечения, целью которой является повышение эффективности проектирования за счет автоматизации семантического анализа текста технического задания. Задачей исследования является автоматизированная обработка текста документа, который изначально создается человеком на естественном языке с соблюдением структуры, требования к которой изложены в ГОСТах.

Предлагаемая методика анализа текста технического задания (рис. 1) состоит из трех этапов: семантическая обработка текста, создание фреймовой структуры и создание диаграмм потоков данных системы, описанной в техническом задании.

Для реализации первого этапа методики была разработана семантическая модель текста технического задания, включающая требования, сформулированные в виде документа на ограниченном естественном языке; второго этапа - фреймовая структура, являющаяся внутренним представлением требований; третьего этапа - модель программного обеспечения в виде описания требований на графическом языке Data Flow Diagrams.

Общий алгоритм семантического анализа текста технического задания состоит из следующих блоков: предварительная обработки текста, синтаксический и семантический анализ и построение модели программного обеспечения.

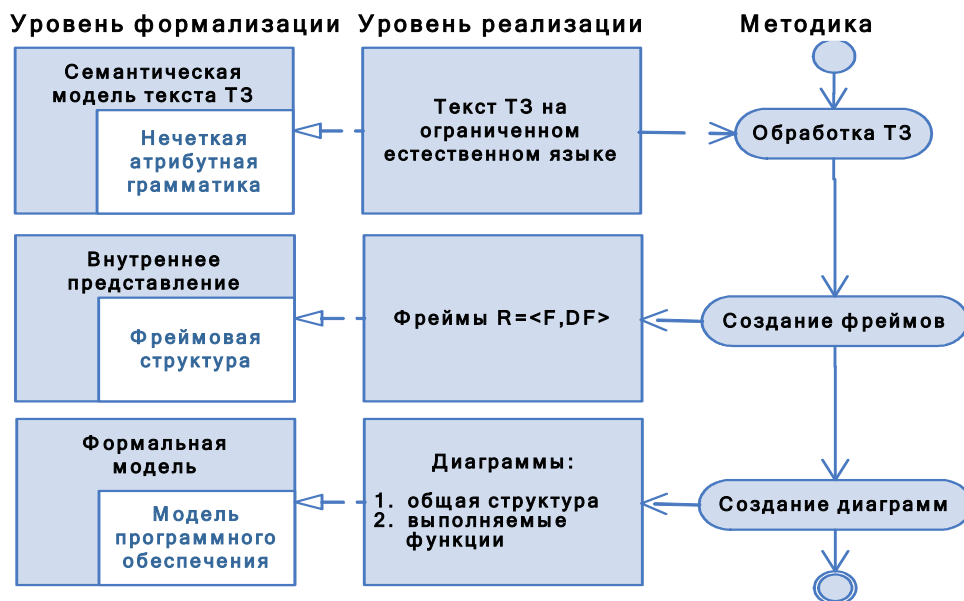


Рисунок 1- Методика анализа текста технического задания

Предварительная обработка текста осуществляется с использованием аппарата конечных автоматов, один из которых представлен на рисунке 2.

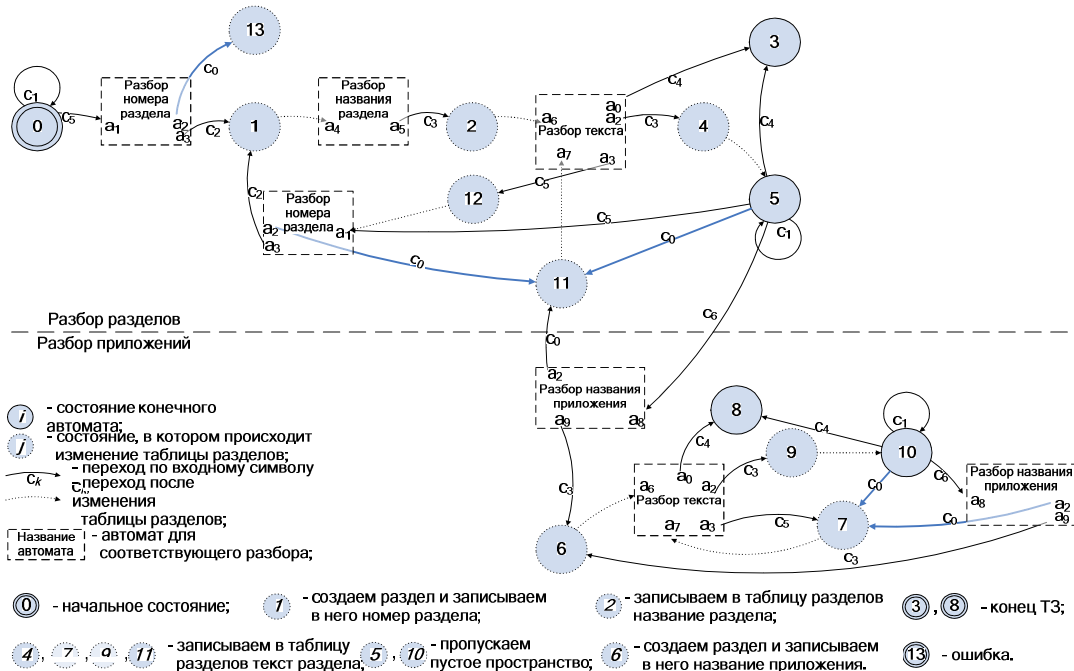


Рисунок 2 - Автомата разбора верхнего уровня текста технического задания

Входные символы конечного автомата: c_1 - пустое пространство, c_2 - пробел, c_3 - новая строка, c_4 - конец текста, c_5 - '1'..'9', c_6 - 'П', c_0 - любой другой символ. Промежуточные состояния автомата: a_1 - начало разбора номера раздела, a_2 - последовательность символов – текст, a_3 - последовательность символов – нумерация, a_4 - начало разбора названия раздела, a_5 - последовательность символов – название раздела, a_6 - начало разбора текста раздела или приложения, a_7 - последовательность символов – продолжение текста раздела или приложения, a_8 - начало разбора названия приложения, a_9 - последовательность символов – название приложения, a_0 - конец ТЗ.

В ходе работы конечного автомата символы, поступающие на его вход, накапливаются в буфере. В определенных состояниях конечного автомата осуществляется запись текущего содержимого буфера в одну из таблиц, после чего буфер опустошается. Работа автомата продолжается до достижения конечного состояния.

На выходе алгоритма предварительной обработки текста формируется набор таблиц: разделов, предложений и лексем. После этого полученные таблицы поступают на вход алгоритма семантического анализа.

Семантическая модель текста технического задания содержит разработанную расширенную нечеткую атрибутивную грамматику над фреймовой структурой формального документа “Техническое задание”, которая позволяет наиболее полно отобразить содержимое ТЗ.

Расширенная нечеткая атрибутивная грамматика, определена в виде:

$$AG = \langle N, T, P, S, B, F, A, D(A) \rangle. \quad (1)$$

где N – конечное множество нетерминальных символов; T – непересекающееся с N множество терминальных символов; P – конечное множество правил; S – выделенный символ из N , называемый начальным символом; B – множество лингвистических переменных $\beta_{k,i}$, соответствующих терминальным символам T (переменная i на k уровне); F – множество функций принадлежности $f_{k,i}$, определяющих степень принадлежности $\mu_{k,i}$ лингвистических переменных; A – множество атрибутов, $A = A_{\text{sin}} \cup A_{\text{sem}}$, где A_{sin} – синтаксические атрибуты, A_{sem} – семантические атрибуты; $D(A)$ – конечное множество семантических действий.

Лингвистические переменные из множества $B = \{\beta_{k,i}\}_{k,i}$, используемые для анализа текста технического задания, описываются:

$$\beta_{k,i} = \langle \beta, T(\beta), U, G, M \rangle, \quad (2)$$

β – название лингвистической переменной (наименование и область применения, основание для разработки, технические требования к программному изделию, стадии и этапы разработки и т.д.);

$T(\beta)$ – языковые выражения. Для лингвистических переменных верхнего уровня они являются лингвистическими переменными, соответствующими терминалам правой части правила. Для лингвистических переменных нижнего уровня – нечеткими переменными, то есть выражениями естественного языка.

U – универсум, $T(\beta) \subset U$;

G – правила морфологического и синтаксического описания языковых выражений, которые определяют синтаксические атрибуты A_{sin} ;

M – семантическое правило для лингвистических переменных, которое индуцируется морфологическими и синтаксическими правилами, так как смысл термина в T частично определяется его синтаксическим деревом, и семантическими атрибутами A_{sem} .

Язык представления выражений состоит из констант и правил их последовательного применения. На морфологическом уровне константами являются граммы (рд – родительный падеж, мн – множественное число). На синтаксическом – названия отношений и групп (ПОДЛ – отношение между подлежащим и сказуемым, ПГ – предложная группа). Для каждого слова входного текста выдается множество морфологических интерпретаций следующего вида: лемма; морфологическая часть речи; набор общих грамм; множество наборов грамм. Методы представления связей между правилами транслируются на язык нечеткой математики. При этом связи представляются нечеткими отношениями, предикатами и правилами, а последовательность преобразований этих отношений – как процесс нечеткого вывода.

Лингвистические переменные верхнего уровня являются составными, то есть включают лингвистические переменные нижнего уровня. Благодаря этому можно построить дерево лингвистических переменных и установить зависимость между ними.

Функции принадлежности из множества $F=\{f_{k,i}\}_{k,i}$ лингвистических переменных $\{\beta_{k,i}\}_{k,i}$, необходимы для построения нечеткого вывода. В частности, каждому правилу грамматики из множества P ставится в соответствие функция принадлежности $f_{k,i}$. Эта двойственная система подстановок используется для вычисления смысла лингвистической переменной.

Синтаксические атрибуты A_{sin} , используемые в грамматике: 'Название' – текст представляет собой наименование раздела; 'Содержимое' – текст представляет собой содержимое раздела; 'Клауза' – клауза; 'Клауза ТИРЕ'- Фрагмент с тире; 'Группа ГЕНИТ_ИГ' – именительная группа, связанная родительным падежом и др.

Семантические атрибуты, используемые в грамматике, содержат название атрибута A_{sem} и семантическое действие $D(A)$: "Фрейм СИСТЕМА=Создание" – создается фрейм СИСТЕМА; "Слот НАЗВАНИЕ СИСТЕМЫ=Присваивание" – значение присваивается слоту НАЗВАНИЕ СИСТЕМЫ ФУНКЦИЮ и др.

На последнем этапе методики требования представляются в виде набора диаграмм потоков данных на графическом языке Data Flow Diagrams: общая структура системы с указанием ее входных, выходных потоков и функции, выполняемые системой, с их входными, выходными потоками.

Указанные выше этапы реализованы в разработанной авторами автоматизированной системе семантического анализа текста технического задания "Семантика ТЗ", которая состоит из следующих подсистем: "Хранение документов", "Интерфейс", "Предварительная обработка текста", "Синтаксический анализ", "Семантический анализ", "Построение диаграмм потоков данных". Проект разработан на платформе Microsoft.NET Framework (язык разработки C#). Таблицы разделов хранятся в формате XML, а их визуальное представление возможно с использованием XSL-преобразования. Полученное при семантическом анализе фреймовое описание также сохраняется в формате XML. Построение диаграмм потоков данных осуществляется с помощью взаимодействия системы с программой MS Visio.

Научная новизна исследования состоит в следующем: разработана методика анализа текста технического задания на начальных этапах проектирования программного обеспечения, включающая семантическую модель текста технического задания, преобразование смысла текста во фреймовую структуру и построение на ее основе модели программного обеспечения.

1. Впервые семантическая модель текста технического задания представлена как расширенная нечеткая атрибутивная грамматика над фреймовой структурой, содержащая синтаксические, семантические атрибуты и лингвистические переменные.

2. Предложены и разработаны методика, предназначенная для преобразования исходного текста технического задания в разработанную фреймовую структуру, представляющую собой описание требований к проектируемому программному обеспечению.

3. Разработан метод построения модели программного обеспечения, описанного в техническом задании, в виде диаграмм потоков данных.

Работа выполнена при поддержке грантов РФФИ 10-01-00135-а, 10-01-00165-а, 10-01-90012-Бел_а

Библиографический список

[Заболеева-Зотова, 2007] Заболеева-Зотова, А.В. Моделирование лексического анализа текста технического задания / А.В.Заболеева-Зотова, Ю.А.Орлова // Известия ВолгГТУ. Серия "Актуальные проблемы управления, вычислительной техники и информатики в технических системах": межвуз. сб. науч. ст. / ВолгГТУ.- Волгоград, 2007.- Вып. 2, №2.- С.39-42.

[Заболеева-Зотова, 2007] Заболеева-Зотова, А.В. Автоматизация процедур семантического анализа текста технического задания / А.В. Заболеева-Зотова, Ю.А. Орлова // Известия ВолгГТУ. Серия "Актуальные проблемы управления, вычислительной техники и информатики в технических системах": межвуз. сб. науч. ст. / ВолгГТУ. - Волгоград, 2007. - Вып.3, №9. - С. 49-52.

[Заболеева-Зотова, 2008] Заболеева-Зотова, А.В. Атрибутивная грамматика формального документа "Техническое задание" / А.В. Заболеева-Зотова, Ю.А. Орлова // Известия ВолгГТУ. Серия "Актуальные проблемы управления, вычислительной техники и информатики в технических системах": межвуз. сб. науч. ст. / ВолгГТУ. - Волгоград, 2008. - Вып. 4, № 2. - С. 39-43.