УДК 811.322

МАТЕМАТИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ПОДСИСТЕМЫ СИНТЕЗА ТЕКСТА В САПР

И.Л. Коробова

Кафедра «Системы автоматизированного проектирования», ГОУ ВПО «ТГТУ»; ira.sapr.tstu@mail.ru

Представлена членом редколлегии профессором В.И. Коноваловым

Ключевые слова и фразы: грамматика языка; диалоговая система; компьютерная лингвистика; синтаксис; семантика; фрейм; шаблонизация.

Аннотация: Рассмотрена методика разработки систем синтеза текста на естественном языке. Приведено описание математического обеспечения, в том числе методов разработки шаблонов, генерации текста, синтаксического анализа.

При разработке подсистемы синтеза текста [1, 2, 5] мы ставили задачу формализовать процесс шаблонизации [3], сделать его понятным и применимым в различных процедурах САПР. В связи с этим, наиболее трудоемким оказался процесс описания математического обеспечения.

Математическое обеспечение подсистемы автоматизированного синтеза текста на основе технологии шаблонизации состоит из множества методов, образующих три группы: методы разработки шаблонов, методы генерации текста, методы обеспечения синтаксического анализа и корректировки.

I. *Методы разработки шаблонов включают:*

– разработку шаблонов последовательности и связей функций, которые отвечают за правильное следование функций и правильные связи между ними;

– разработку шаблонов функций, которые предусматривают возможность добавления или удаления элемента функции, изменения приоритета элемента, его редактирование;

– разработку шаблонов элементов функций, которые предусматривают возможность добавления или удаления подэлемента, изменения приоритета подэлемента (используется сдвиг набора подэлементов), изменение значения (если это глагол-действие), изменение падежа (если это имя существительное или прилагательное).

II. *Методы генерации текста включают:*

– генерацию текста-шаблона, в котором отсутствуют индивидуальные данные. Производится перебор всех функций набора, начиная с первой. Для каждой функции производится перебор всех элементов в зависимости от приоритета, начиная с наименьшего. Для каждого элемента производится перебор всех подэлементов в зависимости от приоритета, начиная с наименьшего. Для каждого подэлемента производится вывод в текст-шаблон значения подэлемента;

– генерацию текста, в котором присутствуют индивидуальные данные. Подобен предыдущему методу, но в значения подэлементов, характеризующихся

индивидуальными данными, подставляются введенные ранее индивидуальные данные.

III. *Методы обеспечения синтаксического анализа и корректировки*.

Для их реализации в COM-сервер MS WordDocument передается текст, и в диалоговом режиме производится проверка правописания, затем производится ручная корректировка текста – в выходном тексте заменяются выбранные пользователем фрагменты.

Рассмотрим подробно алгоритм блока формирования шаблона последовательности функций для выбранного сюжета (рис. 1).

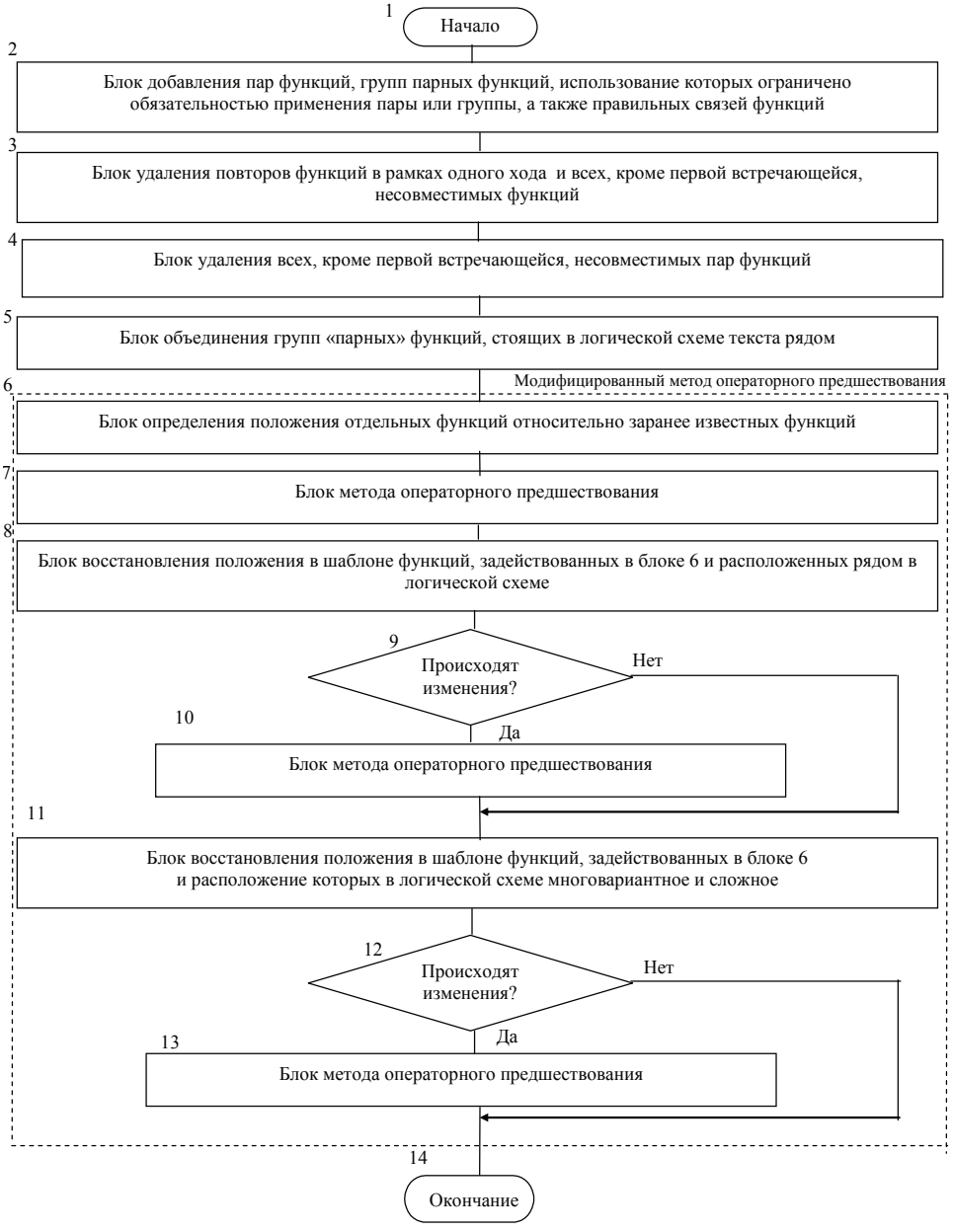

Рис. 1. Блок формирования шаблона последовательности функций для выбранного сюжета

Блок определения наличия в массиве функций, необходимых для существования текста, основан на правиле:

ЕСЛИ $f_j <> f_i$ ТО count[j] := count[j] + 1;
ЕСЛИ count[j] = n
ТО количество несовпадений := количество несовпадений + 1; (1)
ЕСЛИ количество несовпадений > 0, ТО результат := ложь
ИНАЧЕ результат := истина;

где n – количество функций в наборе; count[j] – счетчик несовпадений f_j функции, необходимой для существования текста ($j \in [1, m]$), с f_i функциями набора ($i \in [1, n]$), m – количество функций, необходимых для существования текста; количество несовпадений – количество функций из m, которые не встретились в тексте; результат – наличие в наборе m функций, необходимых для существования текста.

Блок определения сюжета текста основан на правиле:

ЕСЛИ $f_j = f_i$ ТО stroka[j] := 1;
$$\text{Сюжет} := \sum_{j=1}^{m} \left(\text{strto int}(j) 2^j\right) + 1 \quad (2)$$

где stroka[j] – строка (изначально равна ‘000…00’) со строковым номером сюжета в двоичном виде ($j \in [1, m]$); сюжет – номер сюжета в десятичном виде (сюжет $\in [1; 2^m]$).

Блок добавления пар функций, групп парных функций, правильных связей функций основан на правиле:

ЕСЛИ $f_j^l = f_i$ ТО [ЕСЛИ НЕ $f_r <> f_k^l$ ТО count[k] := count[k]+1]
ЕСЛИ count[k] = n ТО [массив[n] := f_k^1; n := n+1] (3)

где массив[n] – ячейка набора функций.

Добавление правильных связей функций выполняется блоком добавления правильных связей функций. Используется утверждение, что не для всех функций набора необходимы взаимные связи.

Блок удаления повторов функций в рамках одного хода и всех, кроме первой встречающейся, несовместимых функций основан на правиле:

ЕСЛИ $f_j = f_i$ ТО [ЕСЛИ $f_r = f_j$ ТО (массив[p] := массив[p + 1]; n := n – 1)] (4)

Блок удаления всех, кроме первой встречающейся, несовместимых функций основан на правиле:

ЕСЛИ $f_j^l = f_i$ ТО [ЕСЛИ $f_r = f_k^l$ ТО (массив[p] := массив[p + 1]; n := n – 1)] (5)

где l – номер группы несовместимых функций ($l \in [1, m]$); s – количество функций в l-й группе ($j \in [1, s]$, $k \in [1, s]$, $j \neq k$); массив[p] – ячейка набора функций.

Блок объединения групп «парных» функций, стоящих в логической схеме текста рядом, основан на правиле:

ЕСЛИ $f_1^l = f_i$ ТО [ЕСЛИ (($f_r = f_k^l$)И ($k <> r$)) ТО
ЕСЛИ $r > i$ ТО (переменная := массив[r];
массив[p] := массив[p – 1];
массив[$i + k + 1$] := переменная) (6)
ИНАЧЕ (переменная := массив[r];
массив[p] := массив[p + 1];
массив[$i + k$ – 1] := переменная)]

Блок определения положения отдельных функций относительно заранее известных функций основан на правиле:

ЕСЛИ $f_i = f_j$ ТО [ЕСЛИ $i < m$ ТО символ := ’<’ ИНАЧЕ символ := ’>’] (7)

где символ – ячейка матрицы, схожей с матрицей операторного предшествования (изначально равен ’ ’).

Блок коррекции порядка следования функций основан на методе операторного предшествования (между функциями f_i и f_j существует отношение ‘<’, если $i < j$; отношение ‘>’, если $i > j$; ‘n’ (несовместимость), если f_i и f_j не могут располагаться рядом в наборе функций.

Модифицированный метод операторного предшествования основан на правиле:

ЕСЛИ (($i > j$ И матрица[i, j] = ’<’) ИЛИ ($i < j$ И матрица[i, j] = ’>’)) ТО
[ЕСЛИ $i > j$ ТО ($k1 := j$; $k2 := i$)
ИНАЧЕ ($k1 := i$; $k2 := j$)
переменная := массив[$k1$]; (8)
массив[p] := массив [$p – 1$];
массив[$k2$] := переменная]
ИНАЧЕ переход на следующий шаг

Блок восстановления положения в шаблоне функций, задействованных в блоке определения положения отдельных функций относительно заранее известных функций, расположенных рядом в логической схеме, основан на правиле:

ЕСЛИ $f_i = f_j$ ТО [ЕСЛИ (($i > m$ И символ = ’<’)
ИЛИ ($i < m$ И символ = ’>’)) ТО
(a := массив[i]; (9)
массив[i] := массив[m];
массив[m] := a)]

Математические методы и алгоритмы синтеза текста решают задачу выбора выходного параметра системы высказываний (1) – (9) на основе правила Modus Ponens [4].

В настоящее время подсистема синтеза текста работает для формирования системы объяснений в процессе принятия решения в условиях нечеткой экспертной информации [4], а также, для создания технического задания на разработку САПР и ее составных частей [5] .

Список литературы

1. Коробова, И.Л. Автоматизированная система синтеза текста на основе технологии шаблонизации [Электронный ресурс] / И.Л. Коробова // Мат. межрегионал. науч.-практ. конф. «Информатизация системы образования Тамбовского региона». – Режим доступа : http://club-edu.tambov.ru/main/news/index.php?r=konf1&f=t12. – Загл. с экрана.

2. Коробова, И.Л. Информационное обеспечение подсистемы синтеза текста при автоматизированном проектировании технологических объектов / И.Л. Коробова, И.А. Дьяков // Теплофизика в энергосбережении и управлении качеством : мат. Шестой междунар. теплофиз. шк. / Тамб. гос. техн. ун-т. – Тамбов, 2007. – Ч. 2. – С. 27–31.

3. Информатика : энциклопед. слов. для начинающих / сост. Д.А. Поспелов. – М. : Педагогика-Пресс, 1994 – 352 с.

4. Коробова, И.Л. Анализ знаний в экспертной системе нечеткого принятия решений / И.Л. Коробова // Вест. Тамб. гос. техн. ун-та. – 2005. – Т. 11, № 4. – С. 873–881.

5. Коробова, И.Л. Подсистема синтеза текста в САПР / И.Л. Коробова, Н.В. Майстренко // Вест. Тамб. гос. техн. ун-та. – 2009. – Т. 15, № 1. – С. 49–55.

Mathematical Software for Subsystem of CAD Text Synthesis

I.L. Korobova

Department "Computer Aided Design Systems", TSTU;
ira.sapr.tstu@mail.ru

Key words and phrases: dialogue system; computer linguistics; grammar of the language; frame; syntax; semantics; standardization.

Abstract: The paper studies the method of developing systems of text synthesis in the natural language. The description of mathematical software, including the techniques for developing templates, text generation and syntactic analysis is given.

Matematische Versorgung des Subsystemes der Synthese des Textes in SAPR

Zusammenfassung: Es wird die Methodik der Erarbeitung der Systeme der Synthese des Textes auf der Natursprache betrachtet. Es ist die Beschreibung der matematischen Versorgung, darunter der Methoden der Erarbeitung der Schablonen, der Textgeneration, der syntaktischen Analyse, angeführt.

Le logitiel du sous-système de la synthèse du texte dans CAO

Résumé: Est examinée la méthode de l'élaboration des systèmes de la synthèse du texte en langue naturelle. Est donnée la description du logitiel y compris des méthodes de l'élaborations des clichés, de la génération des textes, de l'analyse syntaxique.

Автор: ***Коробова Ирина Львовна*** – кандидат технических наук, доцент кафедры «Системы автоматизированного проектирования», ГОУ ВПО «ТГТУ».

Рецензент: ***Подольский Владимир Ефимович*** – доктор технических наук, профессор, заведующий кафедрой «Системы автоматизированного проектирования», проректор по информатизации, ГОУ ВПО «ТГТУ».