УДК 811.322

ПОДСИСТЕМА СИНТЕЗА ТЕКСТА В САПР

И.Л. Коробова, Н.В. Майстренко

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

Представлена членом редколлегии профессором Ю.Л. Муромцевым

Ключевые слова и фразы: диалоговая система; компьютерная лингвистика; фрейм; шаблонизация.

Аннотация: Рассматриваются вопросы разработки систем синтеза текста на естественном языке. Приведено описание проектирующих и обслуживающих подсистем. Показана схема работы подсистемы.

В современных системах автоматизированного проектирования (**САПР**) подсистема синтеза текста имеет важное значение. Она может использоваться при создании диалоговых процедур общения между системой и человеком–специалистом для создания обоснованного технического задания; формирования системы объяснений в процессе принятия решений; формирования проектной документации и т.д.

Для того чтобы диалоговая система могла успешно функционировать, необходимо решить три основные задачи [1]:

1) проанализировать заданный вопрос, выявить его грамматическую структуру, формализовать ее, приведя к типовой форме, доступной восприятию компьютера. Эта задача решается с помощью специальных программ, осуществляющих лингвистический анализ входного текста (вопроса) и выделяющих объекты и отношения между ними, которые позволяют установить, какую информацию следует искать в памяти компьютера;

2) найти среди хранящейся в компьютере информации объекты, указанные в вопросе, и отношения между ними. Данная задача зависит от формы представления информации о рассматриваемой предметной области в компьютере. В простейшем случае вся информация представляется явно, и поиск ответа заключается лишь в сравнении наименований объектов, указанных в вопросе, с теми, которые хранятся в машине. В более сложном случае на основании хранящейся информации формируется модель предметной области, которая используется для получения ответа;

3) преобразовать найденные данные в текст (синтезировать ответ) на естественном языке, согласованный с заданным вопросом. Программа, осуществляющая синтез текста ответа, должна на основе анализа вопроса выбрать грамматически правильную структуру ответа, оценить морфологические особенности входящих в ответ слов, их тип, род, число, время и т.п. И на основании этого преобразовать их так, чтобы они составили грамматически правильный, согласованный текст.

Один из возможных путей синтеза текста состоит в использовании актантов действий. С каждым действием связан некоторый набор соответствующих ему объектов и характеристик. Они, как правило, совпадают с глубинными падежами Филмора. Если, например, мы имеем дело с действием «идти», то с ним тесно

связаны: субъект, совершающий это действие; пункты начала и конца движения; цель движения и т.п. Это позволяет связать с глаголом «идти» некоторую структуру с набором пустых пока мест (рис. 1).

Такие структуры названы фреймами. Заглавными буквами в этой структуре обозначены некоторые имена. Первое имя конкретизируется глаголом «идти», а остальные имена пока остаются незаполненными. Эти остальные имена и определяют актанты глагола «идти».

Наличие актантных структур действий позволяет представить процесс синтеза текстов в виде ряда следующих друг за другом шагов. На первом шаге генерируется нужная последовательность глаголов-действий. На следующем шаге заполняются их актантные структуры, что приводит к появлению глубинной семантической структуры отдельных предложений. Затем эти структуры связываются с учетом общих действующих субъектов и используемых объектов, а также иных связывающих параметров в единый текст. Последний шаг – образование синтаксически правильных конструкций в предложениях.

В разработанной подсистеме для синтеза текста используется технология шаблонизации [2], устанавливающая правила создания и оформления шаблона – некоторого макета текста, определяющего внешний вид данных, но не сами данные. На рис. 2 показано, что использование системы шаблонизации позволяет исключить редактора из процесса программирования и оформления текста, а дизайнера – из процесса программирования и создания текста. Данные хранятся отдельно от программ и шаблонов текста. Например, их можно держать в базе данных и организовать к ним специальный интерфейс для редактирования – backend. Основная его функция – осуществлять легкий доступ к сложным системам хранения данных. Иными словами, редактор не задумывается о том, как будет выглядеть текст и где он будет храниться. Дизайнер же не знает, где хранятся шаблоны. Он просто заполняет форму и сохраняет ее. После этого backend для дизайнера сам проверяет шаблон на правильность и сохраняет его в нужное место.

Подобный подход реализуется в различных разработках [3–5]. Система автоматизированного синтеза волшебных сказок в диалоговом режиме из базы знаний выбирает действующих лиц: «Герой», «Антигерой», «Кого обидели», «Предмет спора», и проводит генерацию сказки [3]. Система Technical Guide Builder, разработанная компанией НИЦ CALS-технологий «Прикладная Логистика», предназначена для автоматизированной подготовки сопроводительной документации на сложные изделия в электронном виде. Данная система осуществляет:

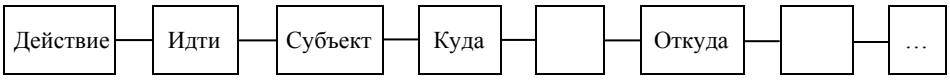

Рис. 1. Пример актантной структуры

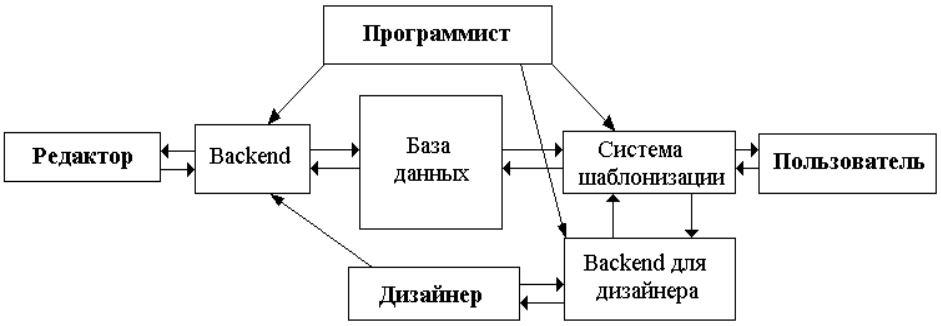

Рис. 2. Схема взаимодействия разработчиков документа на естественном языке

подготовку документации в соответствии с международными стандартами; автоматизированное формирование логических связей между частями и разделами документации; автоматизированное кодирование разделов документации и изделий в электронных каталогах в соответствии с выбранным стандартом; автоматизированный ввод исходных данных из офисных приложений; централизованное управление базой данных проектов документации на изделия [4]. Главное назначением аналитической системы Норильского комбината, разработанного специалистами фирмы Cognitive Technologies, состоит в повышении эффективности экспертной деятельности, увеличении полноты и качества собираемой в Internet-сети информации. Система существенно облегчает и ускоряет работу сотрудников компании, отвечающих за подготовку новостных материалов для руководства. Эта деятельность в принципе типична для любого крупного предприятия. И хотя, конечно, полностью формирование документа эксперт завершает вручную, вся информация для него уже автоматически выбрана, аккуратно рассортирована и структурирована самой системой [5]. В приведенных примерах задача синтеза текста привязана к определенной тематике. В своей работе мы ставим задачу формализовать процесс шаблонизации, сделать его понятным и применимым в различных процедурах САПР.

Составными структурными частями подсистемы автоматизированного синтеза текста на основе технологии шаблонизации являются проектирующие и обслуживающие подсистемы [3].

Проектирующие подсистемы включают:

– разработку интерфейса, отвечающую за проработку задания на проектирование, формирование базы знаний, включающей схемы сюжетов текста, функции, примеры, связи между функциями и их примерами;

– разработку шаблонов, которая в диалоговом режиме производит формирование шаблона последовательности функций, шаблонов отдельных функций, элементов функций. Результатом работы подсистемы является текст, не содержащий индивидуальных исходных данных;

– ввод и редактирование индивидуальных исходных данных текста, которые в диалоговом режиме формируют текст, содержащий индивидуальные данные.

Обслуживающие подсистемы включают:

– управление формированием, поиском, выдачей и корректировкой данных, поступающих от других подсистем запросов и данных;

– синтаксический анализ, отвечающий за формирование корректных окончаний слов текста, содержащего индивидуальные данные;

– вывод проектной документации.

Работа подсистемы осуществляется в диалоговом режиме (рис. 3).

Первоначально пользователь производит осмотр и анализ готовых проектов. При наличии готового решения осуществляется переход на завершающий блок вывода проектной документации.

Для создания нового проекта пользователь разрабатывает шаблон последовательностей функций, шаблоны выбранных функций и элементов функций. Происходит заполнение актантных структур элементов шаблонов функций, то есть ввод индивидуальных данных.

Введение в схему работы блока разработки последовательности функций (блок 5) позволяет сделать подсистему гибкой и применимой для различных приложений. Этот блок определяет сюжет текста (то есть конкретный вид универсальной логической схемы текста), а затем и его содержание как набора функций для заданного сюжета.

Каждая функция состоит из набора элементов, характеризующихся названием, приоритетом (положением в предложении), возможностью редактирования. При разработке шаблона функции (блок 6) предусмотрена возможность добавления или удаления элемента, изменения приоритета элемента (используется сдвиг

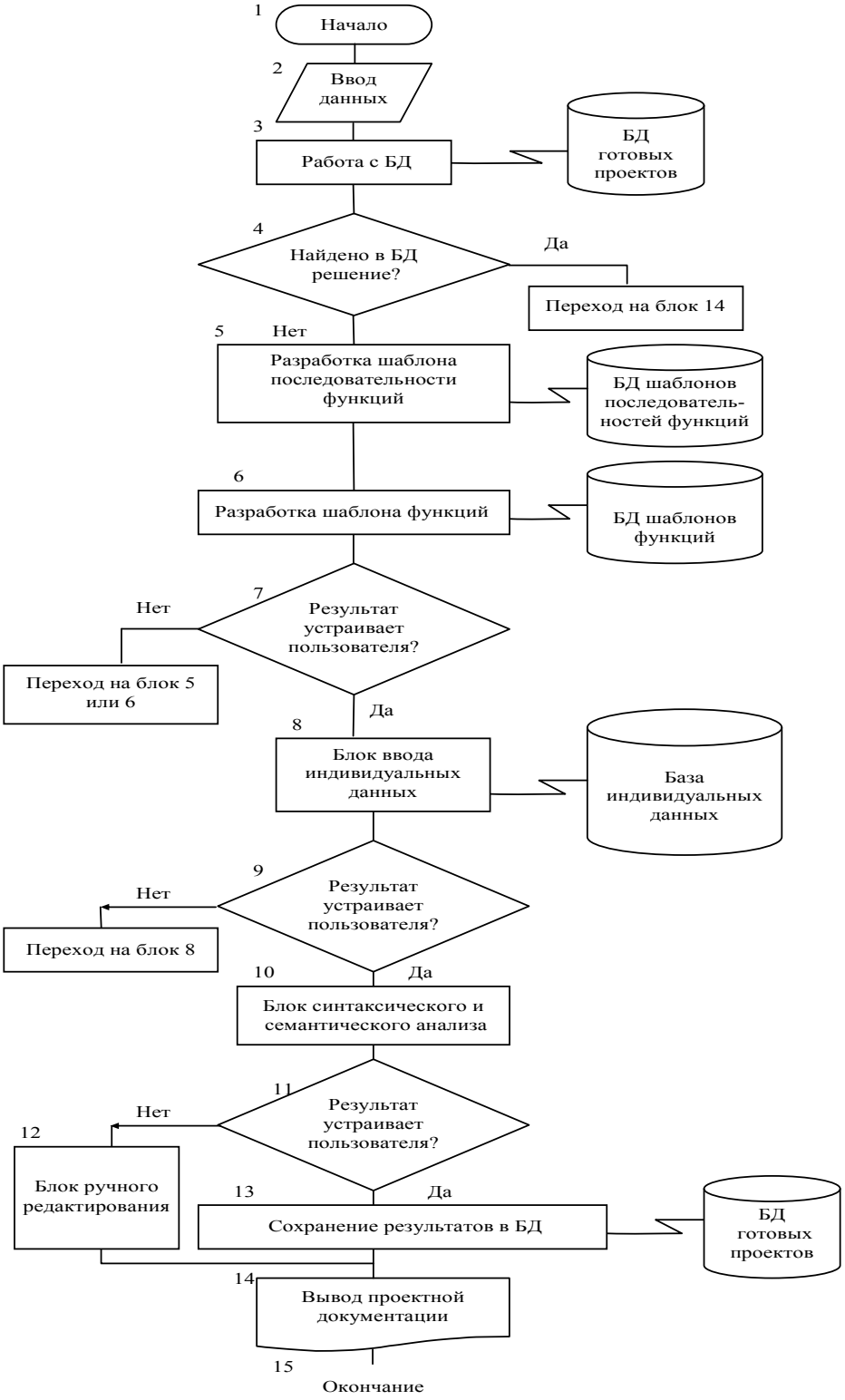

Рис. 3. Схема работы системы

набора элементов), его редактирование (если возможно). В результате формируется промежуточный текст, подвергшийся глубинному семантическому и синтаксическому анализам [4], в котором не достает индивидуальных данных. Если полученный результат не устраивает пользователя, то работа системы возвращается на блок 5 или 6 по выбору.

Заполнение шаблона текста индивидуальными данными (блок 8) осуществляется следующим образом: из таблиц базы данных [5] выбираются определенные индивидуальные данные и включаются в структуру текста (шаблон без индивидуальных данных) в соответствии с математическими алгоритмами.

Для выполнения синтаксического и семантического анализов полученного текста (блок 10) в COM-сервер MS Word Document передается текст с индивидуальными данными, и производится проверка правописания.

Подсистема САПР выполняет сохранение результирующего текста в базу данных (блок 13). Производится вывод проектной документации (блок 14).

Основным достоинством разработанной подсистемы САПР является возможность автоматизированного изменения структуры текста. В настоящее время данная подсистема работает для создания технического задания (**ТЗ**) на разработку САПР, а также для формирования системы объяснений в процессе принятия решения в условиях нечеткой экспертной информации [6].

Подсистема синтеза текста обеспечивает легкое создание профессионального ТЗ на разрабатываемый проект в соответствии с ГОСТ, с возможным последующим редактированием частей проекта на стадиях согласования, возможностью редактирования ранее созданного проекта, экспортом результатов в формате HTML и Microsoft Word.

При создании нового технического задания необходимо выбрать директорию, в которую оно будет сохранено. Автоматически будет создан шаблон готового технического задания, пользователю остается внести изменения в данный шаблон (название проекта, область применения, требования к функциональным характеристикам, надежности, условия эксплуатации и т.д.) (рис. 4). На усмотре-

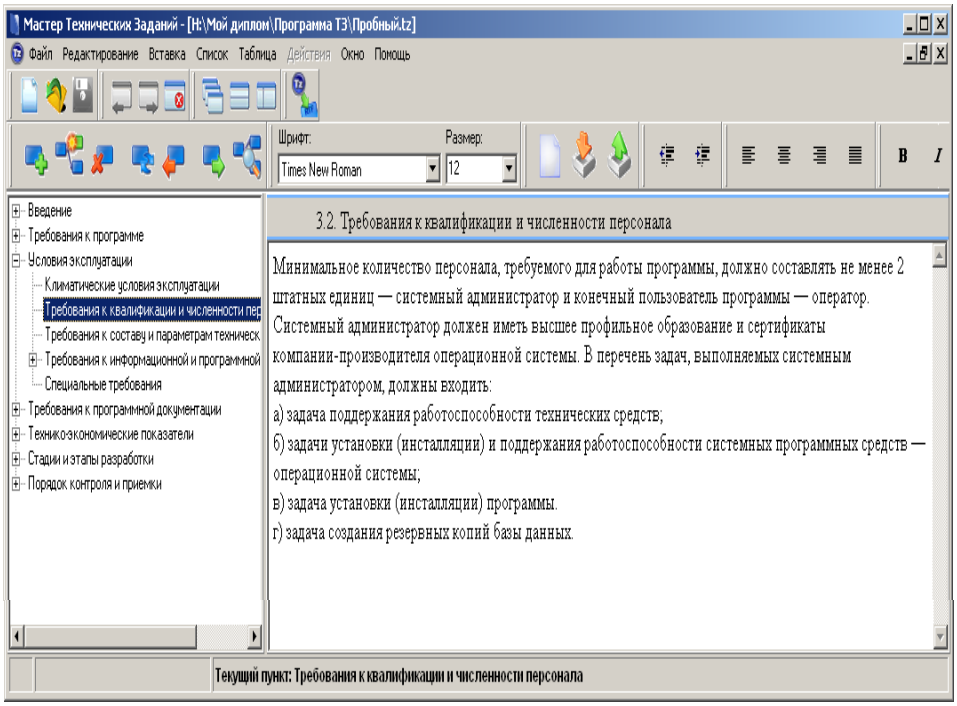

Рис. 4. Работа процедуры составления ТЗ

ние составителя ТЗ в него можно добавить дополнительные пункты, удалить пункты, которые не нужны, изменить шрифт, добавить необходимые рисунки. После того как будет составлено ТЗ, его можно предварительно посмотреть, без экспорта в HTML или Microsoft Word, исправить необходимые пункты, вывести на печать или экспортировать. Программа проста в использовании и не требует подключения дополнительных модулей и библиотек.

Список литературы

1. Информатика: Энциклопедический словарь для начинающих / сост. Д.А. Поспелов. – М. : Педагогика-Пресс, 1994 – 352 с.

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

3. Справочник по САПР / А.П. Будя [и др.] ; под ред. В.И. Скурихина. – Киев : Техника, 1988 – 375 с.

4. Системное программирование. Основы построения трансляторов : учеб. пособие для высш. и сред. учеб. заведений. – СПб. : КОРОНАпринт, 2000. – 256 с.

5. 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: Материалы Шестой междунар. теплофиз. шк. Ч. 2. – Тамбов, 2007. – С. 27–31.

6. Коробова, И.Л. Анализ знаний в экспертной системе нечеткого принятия решений / И.Л. Коробова // Вестн. ТГТУ. – 2005. – Т. 11, № 4. – С. 873–881.

Subsystem of CAD Text Synthesis

I.L. Korobova, N.V. Maistrenko

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

Key words and phrases: computer linguistics; dialogues system; frame; tinplating.

Abstract: The paper studies the matters of designing the text synthesis systems in natural language. The description of designing and servicing subsystems is given. The subsystem flow chart is shown.

References

1. Information Science : Encyclopaedic dictionary for beginner / author: D.A. Pospelov. – M. : Pedagogika-Press, 1994. – 352 p.

2. Korobova, I.L. Automatized system of text synthesis on mould technology / I.L. Korobova // Materials of Regional Scientific-Practical Conference "Information Science Systems of Education of Tambov Region". – URL : http://club-edu.tambov.ru/main/news/index.php?r=konf1&f=t12.

3. Reference book of CAD / A.P. Budya [et al.]. – Kiev : Technique, 1988. – 375 p.

4. System programming. Foundations of construction translators : Educational textbook. – SPb. : KORONA-Print, 2000. – 256 p.

5. Korobova, I.L. Information security of subsystem of text synthesis by automat zed designing of technological object / I.L. Korobova, I.A. Dyakov // Materials of Sixth International Heat-Physics school. Part 2. – Tambov, 2007. – P. 27–31.

6. Korobova, I.L. Analysis of knowledge in expert system of fuzzy decision-Making / I.L. Korobova // Transactions TSTU. – 2005. – V. 11, № 4. – P. 873–881.

Subsystem der Textsynthese im SAPR

Zusammenfassung: Es werden die Fragen der Erarbeitung der Systeme der Textsynthese auf der Natursprache betrachtet. Es ist die Beschreibung der projektierenden und bedienenden Subsysteme angeführt. Es ist das Schema des Subsystemfunktionierens angezeigt.

Le sous-système de la synthèse du texte en CAO

Résumé: Sont examinés les problèmes de l'élaboration des systèmes de la synthèse du texte en langue naturelle. Est citée la comparaison par pairs des descriptions des sous-systèmes de conception et de service. Est montré le schéma du fonctionnement du sous-système.

Авторы: ***Коробова Ирина Львовна*** – кандидат технических наук, доцент кафедры «Системы автоматизированного проектирования»; ***Майстренко Наталья Владимировна*** – кандидат технических наук, доцент кафедры «Системы автоматизированного проектирования», ГОУ ВПО «ТГТУ».

Рецензент ***Литовка Юрий Владимирович*** – доктор технических наук, профессор кафедры «Системы автоматизированного проектирования» ГОУ ВПО «ТГТУ».