УДК 651.326

Б.Н. СУДАКОВ, канд. техн. наук, проф., НТУ "ХПИ", Харьков,
А.С. МАЛЁНКИН, магистр, НТУ "ХПИ", Харьков

МЕТОДЫ СИНТЕЗА ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ В ЭКСПЕРТНЫХ СИСТЕМАХ

Был проведен анализ методов синтеза естественно-языковых текстов в экспертных системах, определены и проанализированы основные этапы синтеза, а также определена целесообразность использования того или иного метода в различных задачах. Рассмотрены проблемы, которые встречаются при семантическом синтезе естественно-языковых текстов. Библиогр.: 10 назв.

Ключевые слова: естественно-языковой текст, анализ, синтез, экспертные системы.

Постановка проблемы и анализ литературы. Методы синтеза естественно-языковых текстов являются важной составляющей при построении интерфейса человек-компьютер, который основан на лингвистическом процессоре. Вследствие чего представляется целесообразным провести анализ существующих методов синтеза, для дальнейшего использования их на практике.

В [1] описываются механизмы работы лингвистического процессора. Классификация интеллектуальных систем по структуре и по решаемым задачам представлена в [2, 3]. В [4] представлена проблема создания систем, предоставляющих человеку возможность общаться с вычислительными машинами на естественном языке (в частности, на русском). Модели хранения знаний, а также методы работы с ними представлены в [5]. В [6] рассмотрены экспертные системы и примеры их реализации. Приведены варианты реализации основных блоков экспертной системы. Новые тенденции и прикладные аспекты инженерии знаний, а также программный инструментарий разработки систем, основанных на знаниях, изложен в [7, 8]. В [9] рассматриваются как традиционные в лингвистике подходы к описанию естественного языка и его моделирования средствами вычислительной техники, так и результаты исследований, проведенных в последнее время. В [10] излагаются основные вопросы семантики как раздела курса общей теории языка.

Цель статьи – анализ существующих методов синтеза естественно языковых текстов, а также разрешение проблем, которые возникают в процессе синтеза текста.

Попытки формализовать интеллектуальную деятельность человека привели к постановке фундаментальной лингвистической задачи, состоящей в моделировании его языкового поведения, т.е. в построении функциональной модели естественного языка. Естественный язык служит человеку для выражения собственных мыслей и для понимания мыслей других людей. Формальные модели языка рассматриваются как компоненты различных прикладных ЕЯ-систем. Компонента системы, реализующая формальную лингвистическую модель и способная работать с ЕЯ во всем его объеме, называется лингвистическим процессором (ЛП) [1]. Две основные функции ЛП состоят в извлечении смысла из заданного текста и в выражении заданного смысла текстом на ЕЯ, иначе это функции [2]: моделирования понимания (анализ); моделирования производства текстов (синтез).

В настоящее время письменный ЕЯ текст широко распространен как средство коммуникации пользователя с компьютерными системами. Большая часть программных приложений, которые выдают тексты на ЕЯ, работает с ЕЯ информацией как со строкой символов. Они манипулируют готовыми предложениями и словосочетаниями как строительными блоками будущего текста. Это – шаблонные технологии. Шаблонные технологии относительно просты и надежны и находят широкое промышленное применение. Главная особенность этих технологий состоит в том, что содержание будущего текста представлено в них в виде фрагментов текста.

Другой вид систем работает с содержанием будущего текста, представленном в виде данных нетекстовой природы (БД, баз знаний, семантических и формальных языков) [3]. В этом случае для создания текста системе необходимы знания структуры содержания и знания об устройстве генерируемого текста, а также сложные лингвистические знания, которые позволяют выразить это содержание языковыми средствами.

Шаблонные системы. Шаблонная система использует готовые реплики или комбинирует готовые фрагменты текста таким образом, что они занимают заданные позиции в дискурсе или стереотипном тексте. Самые простые шаблонные системы просто вставляют фрагменты текста в шаблоны без их дополнительной обработки, например, реплика системы: "Не могу найти myl.txt, my2.txt файл(ы)!".

Более сложные шаблонные системы дополнительно проводят ограниченную лингвистическую и риторическую обработку результата – позволяют задавать отдельные грамматические параметры текста или комбинировать шаблонные высказывания в связный текст, используя определенные лексические и грамматические знания о ЕЯ.

Системы автоматической генерации текстов на ЕЯ (ГЕЯ). Генерация на естественном языке (ГЕЯ) как научное направление занимается созданием компьютерных систем, производящих тексты на естественном языке (ЕЯ) из некоторого нелингвистического (нетекстового) представления информации.

ГЕЯ активно развивается за рубежом, начиная с 90-х годов.

JIM системы предназначены для создания текстов, имеющих относительно свободное содержание, которое не может быть заранее задано в виде готовых фрагментов текста. Источником содержания являются данные, представленные в виде БД, БЗ или в виде выражений на формализованных языках, например, SQL.

Несмотря на сложность JIM систем, они имеют важные преимущества по сравнению с использованием ручного труда и шаблонными системами [4]. Лучшее качество создаваемых текстов, многоязыковой выход и гарантированное соответствие стандартам.

Организация системы ГЕЯ: ресурсы и обрабатывающий компонент. Программный код, реализующий генератор обычно разделяется на два компонента: ресурсы и обрабатывающий компонент. Ресурсы описывают знания, необходимые для генерации, а обрабатывающий компонент применяет эти знания к входным данным (словари).

Ресурсы должны быть максимально независимы от особенностей конкретного приложения, а все особенности задачи должны быть отражены в принимаемых генератором на входе данных [5]. В этом случае задачу построения генератора необходимо решить только один раз для каждого описываемого ЕЯ. К настоящему моменту считается, что генератор, реализующий такую схему, на входе должен принимать систему знаний, из которой будет конструироваться текст, коммуникативную цель порождаемого текста, модель адресата текста, и контекст повествования.

Общая и полная схема генерации без детализации происходящих процессов состоит из трех основных блоков:

– планирование содержания текста;

– микропланирование;

– реализация на ЕЯ.

Планирование содержания текста – решение, какая именно информация из входных данных попадет в текст, и как она будет организована [6].

Микропланирование – это интерфейсный блок, который позволяет от предметных знаний перейти к языковым. В нем решается, каким образом выбранная информация будет реализована языковыми

средствами в виде предложений на ЕЯ. Результатом этого процесса являются представления предложений в виде структур семантических и/или синтаксических отношений.

Реализация на ЕЯ – производство грамматически правильных предложений текста [7]. Основано на лингвистических знаниях. Этот блок часто выделяется как универсальный и включает в себя либо только морфологический синтез словоформ, либо переход от семантических представлений к поверхностно синтаксическим и синтез словоформ.

Определение вида входных данных является кардинальным вопросом для JIM систем. Теоретическое рассмотрение возможных типов входов оказывается не эффективным [8]. Можно рассматривать три вида возможных входов для систем ГЕЯ – числовые данные, структурированные объекты и логические формулы. Особенность практических систем состоит в том, что они обычно используют в качестве входа представления данных, порожденные другими системами для некоторых практических целей, а не созданные вручную разработчиками системы. Можно выделить три вида таких входов [9]:

1. БД. Особенность этого типа источника состоит в том, что информация не организована для передачи адресату. Тип текста, который можно построить на основе этой информации, и его структура должны быть определены извне.

2. Семантическое представление – представление содержания текста, созданное человеком с помощью системы интерфейсного типа "человек – компьютер".

3. Представление знаний на формальном языке, например, SQL, логические языки.

Можно выделить два вида данных БД, которые рассматриваются как источники информации для JIM систем ГЕЯ: данные, описывающие некоторые объекты и их признаки, например, БД по товарам, однотипным объектам типа военных кораблей, подержанных автомобилей для продажи и т.п. Другой вид – это поток данных, отражающий состояние одного или группы однотипных объектов в некоторые моменты времени, например, метеорологические замеры (погода), статистические данные по занятости населения (занятость) и др.

Поток данных – "простой отчет". Содержание текста выбирается из исходной БД. В него попадает только та информация, которая интересует пользователей отчетов. Выбранная информация упорядочивается локально по тематическому принципу, заданному извне, а подача информации в целом соответствует зафиксированному в БД потоку данных. В простых отчетах могут моделироваться простейшие

анафорические ссылки в виде замены описательной номинации объекта на указательную – личное местоимение [10].

Тексты типа "связный отчет" описывает ситуацию, характеристики которой – различные объекты. Текст создается на основе БД по тому же общему сценарию, что и простой отчет. Различие состоит в том, что дискурс текста представляет собой не просто последовательность тематических блоков, а некоторую структуру, образованную семантическими и концептуальными связями – план текста.

Генерация текстов из семантического представления. Кардинально отличается от БД другой источник содержания текста – семантические представления [3]. Этот вид исходных данных создается человеком в режиме интерфейса с компьютером.

Моделирование структуры текста обычно выполняется для рассматриваемых систем в специальной системе планирования содержания, имеющей вид, в частности, графического редактора. Графический редактор последовательно предоставляет пользователю возможность выбирать понятия из определенных в данной семантической среде, следуя заданной в нем стратегии организации текста. Таким образом, графический редактор контролирует правильность получающегося представления, предлагая пользователю для продолжения структуры текста допустимые по структурным и семантическим свойствам понятия МПО [5]. Особенность систем данного типа состоит в получении в результате планирования содержания целого связного представления текста, структуру которого образуют дискурсные и предметные отношения между пропозициями или высказываниями.

В отличие от генерации текстов отчетов задача микропланирования в ЛМ системах с семантическим входом состоит, как правило, не в добавлении, а в усечении части исходного семантического представления при переходе к средствам ЕЯ [7]. Переход от единой структуры текста к последовательности предложений, а также реализация некоторых синтаксических конструкций предполагает сокращение фрагментов исходного представления, делая его более лаконичным и естественным. Эти процессы получили название агрегация. При агрегации сокращаются дублирующиеся структуры и понятия.

Некоторые виды текстов существуют и используются на некотором формальном языке, отличном от языковой семантики. Это языки математики, например, логические языки, языки спецификаций, например, представления запросов к БД (SQL) [2]. Особенность "текстов" на этих формальных языках состоит в том, что они коммуникативно организованы, т.е. так же, как и тексты на ЕЯ, непосредственно

предназначены для передачи информации. В качестве входа для системы ГЕЯ такие представления получаются в результате работы определенной нелингвистической системы.

Выводы. В результате данной статьи были проанализированы существующие методы синтеза естественно-языкового текста. Были выделены особенности каждого из методов. Показаны преимущества систем автоматической генерации текстов над шаблонными системами.

Список литературы: **1.** *Апресян Ю.Д.* Лингвистический процессор для сложных информационных систем / *Ю.Д. Апресян, И.М. Богуславский, Л.Л. Иомдин и др*. – М.: Наука, 1992. – 416 с. **2.** Искусственный интеллект: В 3 кн. Кн.1. Системы общения и экспертные системы: Справочник / Под ред. Э.В. Попова. – М.: Радио и связь, 1990. – 464 с. **3.** *Евдокимова И.С.* Естественно-языковые системы: курс лекций / *И.С. Евдокимова*. – Улан-Уде: Изд-во ВСГТУ, 2006. – 92 с. **4.** *Попов Э.В.* Общение с ЭВМ на естественном языке / *Э.В. Попов*. – М.: Наука, 1982. – 360 с. **5.** *Елисеева О.Е.* Компьютерная лингвистика, естественно-языковой и речевой інтерфейс: Учебн. пособие. / *О.Е. Елисеева*. – Минск: БГУИР, 2006. – 160 с. **6.** *Балтрашевич В.Э.* Реализация инструментальной экспертной системы / *В.Э. Балтрашевич*. – СПб.: Политехника, 1993. – 238 с. **7.** *Гаврилова Т.А.* Базы знаний интеллектуальных систем / *Т.А. Гаврилова, В.Ф. Хорошевский*. – СПб.: Питер, 2000. – 384 с. **8.** *Джарратано Дж.* Экспертные системы: принципы разработки и программирование / *Дж. Джарратано, Г. Райли*. – М.: "Вильямс". 2007. – 1152 с. **9.** *Шемакин Ю.И.* Начало компьютерной лингвистики / *Ю.И. Шемакин*. – М.: Издательство МГОУ А/О "Росвузнаука", 1992. – 360 с. **10.** *Кобозева И.* Лингвистическая семантика: Учебное пособие / *И. Кобозева*. – М.: Эдичориал УРСС, 2000. – 352 с.

Статью представил д.т.н., проф. НТУ "ХПИ" Серков А.А.

УДК 651.326

Методи синтезу природно-мовних текстів в експертних системах / Судаков Б.М., Мальонкін А.С. // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2012. – № 38. – С. 172 – 177.

Був проведений аналіз методів синтезу природно-мовних текстів в експертних системах, визначено та проаналізовано основні етапи синтезу, а також визначено доцільність використання того чи іншого методу в різних завданнях. Розглянуті проблеми, які зустрічаються при семантичному синтезі природно-мовних текстів. Бібліогр.: 10 назв.

Ключові слова: природно-мовний текст, аналіз, синтез, експертні системи.

UDC 651.326

Methods of synthesis natural language texts in expert systems / Sudakov B.N., Malyonkin A.S. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modeling. – Kharkov: NTU "KhPI". – 2012. – №. 38. – P. 172 – 177.

Methods of synthesis of natural language texts in expert systems were analysed . Have been identified and analyzed the main stages of the synthesis and determined the feasibility of using a particular method in various tasks. Problems encountered in the synthesis of semantic natural language texts was analyzed. Refs.: 10 titles.

Keywords: natural language text, analysis, synthesis, expert systems.

Поступила в редакцию 18.06.2012