УДК 004.048

Б. Н. СУДАКОВ, канд. техн. наук, профессор, НТУ «ХПИ»;
А. А. ШМАКОВА, магистр, НТУ «ХПИ»

СИНТЕЗ СВЯЗНОГО ЕСТЕСТВЕННО-ЯЗЫКОВОГО ТЕКСТА В ЭКСПЕРТНЫХ СИСТЕМАХ

В статье рассматривается проблема наиболее понятного пользователю синтеза естественно-языкового текста применительно к конкретной предметной области и способы ее решения. Для решения данной проблемы в области систем искусственного интеллекта используется несколько компонент синтеза. Статья описывает все компоненты синтеза и возникающие при этом трудности. Определена необходимость использования тех или иных компонент синтеза при генерации ЕЯ связных текстов.

Ключевые слова: синтез ЕЯ текста, морфологический синтез, синтаксический синтез, формальные грамматики, семантический синтез.

Введение. Синтез связного естественно-языкового(ЕЯ) текста является общим направлением искусственного интеллекта и математической лингвистики. В части искусственного интеллекта это означает генерацию грамотного понятного пользователю текста. Решение этой проблемы будет означать создание более удобной формы взаимодействия компьютера и человека.

Анализ последних исследований и литературы. Чтоб полностью взаимодействовать с человеком, экспертная система должна быть полноправным участником диалога. Для этого ей необходимо обладать следующими функциями:

– ведение диалога
– понимание высказывания
– обработка высказываний(анализ)
– генерация текста(синтез)

Компонента синтеза текста на ЕЯ решает в соответствии с результатами, полученными остальными компонентами системы, две основные задачи: генерация смысла, т.е. определение смысла выходного высказывания системы понятном машине, и синтез высказывания, т.е. преобразование смысла в высказывание на естественном языке.

Первая задача является сложной, т.к. тип высказывания зависит от состояния системы и результатов, полученных предыдущими компонентами. Не всегда система может сгенерировать точный ответ. Также при решении задачи формирования смысла выходного высказывания необходимо учитывать прагматический аспект, то есть цели участников общения. Однако в большинстве существующих систем данная задача решается с помощью достаточно простых подходов.

В промышленных системах общения генерация смысла обычно заклю-

чается в редактировании значений атрибутов или выборе шаблона ответа. В экспериментальных системах для выражения смысла генерируется полное семантическое представление, включающее одно или несколько связанных событий (понятий) с одним или несколькими исполнителями на каждую роль.

Вторая задача компоненты синтеза текста состоит в синтезе естественно-языкового выражения, на основе внутреннего представления выходного высказывания. Данная задача подразделяется на этапы семантического, синтаксического и морфологического синтеза. Сложность задачи синтеза определяется требованиями к естественности и выразительной мощности выходных высказываний. В данном случае естественность – это степень близости к естественному языку, то есть наличие таких свойств, как синонимия и омонимия слов и словосочетаний, свободный порядок слов и др. Под выразительной мощностью понимается возможность выразить разнообразные понятия, отношения, кванторы, процедуры и т. п. Естественность и выразительность выходных высказываний в существующих системах могут быть различными. Например, высказывания могут фактически не синтезироваться, а выбираться из заранее подготовленного списка, либо имеется шаблон ответа, в который подставляются некоторые слова, представляющие собой значения искомых атрибутов, при этом может использоваться морфологический синтез. Большая естественность достигается, если выходное высказывание формируется из семантического представления события (или понятия) с применением морфологии, синтаксиса(для определения порядка и формы слов) и семантики(для выбора лексем.

Цель статьи. Рассмотреть и проанализировать компоненты синтеза ЕЯ текста и определить состав системы синтезы для представления конкретной предметной области.

Постановка задачи. Процесс синтеза состоит из определения информации, которая должна быть сообщена пользователю, выделение из общего множества высказываний, интересующих пользователя, разбиение информации на части, соответствующие будущим предложениям, и установление последовательности этих частей, определение лексем и установление последовательности этих частей. За этим решаются задачи связанные с построением синтаксической структуры отельных предложений и приписыванию вершинам структур морфологической информации.

Основная часть. Для решения проблемы синтеза естественно-языкового текста разработано множество систем. ЕЯ системы базируются на использовании морфологической компоненты, некоторые из них так или иначе используют и синтаксическую компоненту. Наиболее развитые и сложные ЕЯ-системы имеют в своем составе также семантическую и прагматическую компоненты и анализируют не только отдельные предложения, но и входной текст в целом.

Целью семантического синтеза является построение модели сюжета,

описанного текстом.

Семантический синтез использует результаты семантического анализа, который выполняет поиск понятий, соответствующих словам, поиск объектов сюжета, построения временной диаграммы.

Для работы семантической компоненты, необходимо создать семантический язык и толково-комбинаторный словарь (ТКС). Под семантическим языком понимается

а) семантический словарь, в который входит словарь элементарных семантических единиц – сем (имен предметов и предикатов), словарь промежуточных семантических единиц и словарь символов, характеризующих коммуникативную организацию смысла: тема – рема, старое – новое, выделено – не выделено и т.п.;

б) правила образования, по которым из семантического словаря могут строится семантические представления высказываний и которые касаются только формальной правильности семантических представлений;

в) правила преобразования, которые задают синонимичность двух семантических представлений.

Кроме того, для использования семантического языка необходимо иметь набор семантических аксиом и набор правил семантической «комбинаторики» – правил расчленения/сочленения семантических представлений при переходе от смысла к тексту и наоборот.

Также словарь должен иметь набор разных значений для одной словесной единицы.

Целью морфологического синтеза является установление морфемного состава слова, а также морфологических принципов, используемых в задачах синтаксического и семантического синтеза.

На вход программы морфологического синтеза поступают лексема в начальной форме и значения свободных грамматических переменных, соответствующих запрашиваемой словоформе данной лексемы или запрос на синтез всех форм заданной лексемы.

Результатом работы программы морфологического синтеза является либо словоформа с запрашиваемыми грамматическими характеристиками, либо все формы заданной лексемы. Морфологический синтез также может оказаться неоднозначным в случае вариативности флексии в какой-либо форме слова или при морфологической омонимии.

Роль синтаксического синтеза заключается в построении простых предложений в определенных границах, описывающих процесс, участников процесса.

Построить синтаксический синтезатор ЕЯ значительно сложнее, чем морфологический по ряду причин: нет достаточно четкой и формальной лингвистической литературы, описывающий какой-либо ЕЯ, грамматика естественного языка принципиально недетерминирована и неоднозначна, синтаксис ЕЯ весьма разнообразен, сложен и произволен (особенно в разговорной

речи и в поэзии). Трудными для автоматической обработки являются такие вполне допустимые в ЕЯ явления, как эллипсис (пропуск обязательных фрагментов предложения в силу возможности их восстановления из предыдущего контекста), парцелляция (разбиение одного грамматического предложения на несколько предложений для усиления акцента на некоторые его фрагменты).

В синтаксическом синтезе используются два вида систем – модульные и интегральные. В системах модульного типа синтаксическая и семантическая компонента разбиты на разные блоки. В системах интегрального типа два этих компонента слиты воедино, поэтому системы интегрального типа работают только в узкой предметной области. Поэтому наиболее эффективными на сегодняшний день являются системы модульного типа, которые направлены на глубокое понимание синтезируемого текста.

Но у модульной системы тоже есть недостатки, например, в вопросе о том, насколько развитым и «семантизированным» должен быть синтаксический этап анализа: это находит отражение в разной степени дифференцированности синтаксических отношений, в разной глубине интерпретации синтаксических отношений, а также в широте привлечения семантической информации при построении синтаксической структуры входного предложения.

Синтаксическая структура предложения может быть представлена в нескольких видах – дерево зависимостей, структуры непосредственных составляющих, ориентированные структуры непосредственно составляющих.

Деревья зависимостей – предложение может быть представлено как линейно упорядоченное множество элементов (словоформ), на котором можно задать ориентированное дерево (узлы – элементы множества).

Каждая дуга, связывающая пару узлов, интерпретируется как подчинительная связь между двумя элементами, направление которой соответствует направлению данной дуги.

Структуры непосредственно составляющих (НС-структуры)— множество отрезков предложения, называемых *составляющими*, которое удовлетворяют следующим условиям:

– в качестве элементов множества отрезков предложения присутствуют само предложение и все его отдельные словоформы;

– в одну составляющую объединяются отрезки непосредственно синтаксически связанные между собой;

– любые две составляющие либо не пересекаются, либо одна из них содержится в другой.

НС-структуры дают возможность выделить в предложении не только отдельные слова, но и некоторые словокомплексы, функционирующие как единое целое. Такие структуры описывают неподчинительные отношения.

Ориентированные структуры непосредственно составляющих (ОНС-структуры) – это структура составляющих, где для каждой неодноэлемент-

ной составляющей определена одна из ее НС в качестве главной (неглавные зависят от главной).

Всякая ОНС-структура имеет свойство определять соответствующее ей дерево зависимостей или НС-структуру (обратное неверно).

ОНС-структуры имеют недостаток, такой же как и деревья зависимостей – неспособность адекватно описывать неподчинительные связи.

Описание синтаксиса ЕЯ может быть описано следующими формальными грамматиками:

Грамматика зависимостей (G_D) (1)

$$G_D = <V_T, V_N, V_S, R_T, R_N>, \quad (1)$$

где V_T – алфавит терминальных символов; V_N – алфавит нетерминальных символов – классов терминалов, V_S – множество корневых классов, $V_S \in V_N$; R_T – множество правил классификации вида $A \rightarrow a$ (терминал a принадлежит классу A); R_N – множество правил кустов вида $A(B_1 \quad B_k * B_{k+1} \quad B_n)$ или A(*), которые для каждого класса A задают его систему управлений (классами B_j), выраженную в терминах классов, с указанием их линейного порядка относительно корня куста и друг друга.

Язык, порождаемый грамматикой зависимостей, – это множество терминальных цепочек $a_1 \ldots a_n$, где каждый символ a_i можно отнести к определенному классу A_i, и для любого A_i в грамматике существует соответствующее правило куста $r \in R_N$.

Контекстно-свободные грамматики (G_{CF})

Вывод каждой цепочки в G_{CF} можно изобразить в виде дерева. Множество поддеревьев дерева соответствует множеству непосредственно составляющих порождаемой цепочки.

Метка корня дерева – название полной составляющей предложения, а метки узлов-сыновей – имена соответствующих непосредственно составляющих.

Ориентированные контекстно-свободные грамматики ($<G_{CF}, \Delta>$).

Δ-ориентировка грамматики G_{CF}, которая вводится следующим образом: из множества правил R выделяется подмножество R^1, в которое входят все правила вида $A \rightarrow \alpha_1 \ldots \alpha_n$ при $n \geq 1$; $\alpha_1, \ldots \alpha_n \in V_G$.

Для каждого из этих правил в цепочке $\alpha_1 \ldots \alpha_n$ маркируется одно из вхождений α_k в качестве главного (например, сверху *). Выделенный элемент может быть как терминальным, так и нетерминальным.

Синтаксическая база данных должна содержать:

– формальное описание грамматики некоторого фиксированного подмножества выбранного ЕЯ;

– описание синтаксических характеристик отдельных лексем или словосочетаний выбранного подмножества ЕЯ (синтаксический класс, синтаксический подкласс, переходность…); все учитываемые синтаксические характеристики могут содержаться в используемой для целей синтаксического анализа морфологической базе данных, в этом случае необходимо иметь про-

граммные средства, позволяющие извлекать их оттуда;

– описание моделей управления лексем выбранного подмножества ЕЯ (при соответствующем выборе метода синтаксического анализа).

Прагматическая компонента системы синтеза определяет отношение человека к тексту, цели высказывания. Факт, передаваемый предложением является пресуппозицией в прагматическом смысле, если говорящий считает ее истинность само собой разумеющейся и исходит из того, что другие участники контекста считают также. Набор пресуппозиций человека определяется на основании тех утверждений, которые он делает, вопросов, которые он задает. Пресуппозиции — это пропозиции, неявно подразумеваемые еще до начала передачи речевой информации. Пресуппозиции, естественно, не обязаны быть истинными.

Выводы. Основной проблемой синтеза ЕЯ является то, что в разных предметных областях одно выражение будет иметь разный смысл. А также в ЕЯ часто используются устоявшиеся словоформы и фразеологизмы, которые трудно синтезировать по заданным правилам.

Для разрешения этой трудности используется сочетание нескольких компонент синтеза, которые отвечают за составление предложения, общего текста в целом и определение смыла этого текста. Таким образом выбираются разные компоненты системы и используемые ими методы, которые лучше всего подходят под синтез текста для определенной экспертной системой.

Список литературы: 1. *Бондаренко М. Ф., Коноплянко З.Д., Четвериков Г.Г.* Основи синтезу над швикодіючих структур мовних систем штучного інтелекту. **2.** *Виноград Т.* Программа, понимающая естественный язык. – М.: Мир, 1976. – 294 с. **3.** *Гладкий А. В.* Формальные грамматики и языки. – М.: Наука, 1973. – 368 с. **4.** *Нелюбин Л. Л.* Компьютерная лингвистика и машинный перевод : Метод. пособие. – М.: ВЦП, 1991. – 151 с. **5.** *Судаков Б.Н., Омар А.Х. Авадала.* Синтез естественно-языковых сообщений при общении пользователя с экспертной системой // Труды II Междунар. науч.-техн. конф. «Электронные информационные ресурсы: проблемы формирования, хранения, обработки, распространения, защиты и использования». – К.: УкрИНТЭИ. – 2001. – С. 14-15.

Поступила в редколлегию 09.10.2012

УДК 004.048

Синтез связного естественно-языкового текста в экспертных системах / Б. Н. Судаков, А. А. Шмакова // Вісник НТУ «ХПІ». Серія: Техніка та електрофізика високих напруг. – Х.: НТУ «ХПІ», 2012. – № 52 (958). – С. 178-184. – Бібліогр.: 5 назв.

У статті розглядається проблема правильного і найбільш зрозумілого користувачеві синтезу природно-мовного тексту стосовно конкретної предметної області й способи її вирішення. Для вирішення даної проблеми в області систем штучного інтелекту використовується декілька компонент синтезу. Стаття описує всі компоненти синтезу і виникаючі при цьому труднощі. Визначено необхідність використання тих чи інших комопнент синтезу при генерації ЕЯ зв'язкових текстів.

Ключові слова: синтез ЕЯ тексту, морфологічний синтез, синтаксичний синтез, формальні граматики, семантичний синтез.

The paper under discussion covers the problem of the correct and most user friendly synthesis of natural language text in relation particular domain and how to resolve it. To solve this problem in the field of artificial intelligence used by multiple components of synthesis. The article describes the synthesis of all the components and the concomitant difficulties. The necessity of using some komopnent synthesis for the generation of NL coherent text.

Key words: Synthesis of NL text, morphological synthesis, syntactic synthesis, formal grammar, semantic synthesis.