OSTIS-2011
(Open Semantic Technologies for Intelligent Systems)

УДК 004.822:514

СИСТЕМА АВТОМАТИЧЕСКОГО ПОРОЖДЕНИЯ АНГЛОЯЗЫЧНЫХ ДЕЛОВЫХ ЭЛЕКТРОННЫХ ПИСЕМ

Т.В. Бусел (*tatsiana-busel@yandex.ru*)
Белорусский государственный лингвистический университет,
г. Минск, Республика Беларусь

В докладе анализируются современные подходы к решению проблемы автоматического порождения текстов, рассматривается созданная система, ориентированная на синтез англоязычных деловых электронных писем: структурно-функциональная схема порождения, лингвистическая база знаний и принципиальный алгоритм, а также дается оценка полученных результатов.

Ключевые слова: деловое электронное письмо, автоматическая обработка текста, логико-семантическая модель текста, лингвистическая база знаний, система автоматического порождения деловых электронных писем.

Введение

В эпоху развития информационных и коммуникационных технологий и их внедрения во все сферы человеческой деятельности проблема автоматического порождения текстов приобрела особую актуальность, в связи с увеличением количества информации, представляемой в виде электронных документов, значительную часть которых составляет деловая корреспонденция на английском языке. Актуальность современных исследований в данной области также связана с активизаций в Республике Беларусь предпринимательской деятельности, расширением и углублением деловых контактов как внутри страны, так и на международном уровне, формированием новых форм деловой коммуникации, вызванных развитием глобальной сети Интернет, необходимостью быстрого ввода, обновления и обмена текстовой информацией в режиме реального времени. Таким образом, обращение к проблеме автоматического порождения англоязычных текстов делового характера диктуется потребностью современного общества в создании компьютерных систем, призванных помочь специалистам, работающим в сфере международных экономических и внешнеторговых отношений, значительно сократить время, необходимое на их создание, снизить вероятность появления ошибок, вызванных недостаточно высоким уровнем владения английским языком и техническими погрешностями. Для решения указанных проблем был разработан и исследован новый тип системы, ориентированной на автоматическое порождения англоязычных деловых электронных писем по выбранной пользователем теме.

1. Современные подходы к созданию систем автоматического порождения текстов на естественном языке

Постановка проблемы автоматического порождения текста и соответственно попытки ее решения с использованием различных подходов предпринимались многими исследователями. История применения вычислительной техники для порождения письменных и устных текстов насчитывает уже более 40 лет и связана с именами таких исследователей, как М.В. Болдасов, А.В. Зубов, Н.Н. Леонтьева, Дж. Лестер, Б.М. Лобанов, М.Г. Мальков, В. Манн, Ю.Н. Марчук, А.С. Нарьяни, Р.Г. Пиотровский, Э. Райтер, Е.Г. Соколова, Э. Хови, М. Эльхадад и др. За эти годы выработаны различные подходы к решению данной проблемы. Результаты исследования

современного состояния дел в области создания интеллектуальных компьютерных систем, свидетельствуют о том, что построение системы порождения может быть осуществлено в рамках двух принципиально различных подходов [Болдасов, 2003; Соколова, 2005; Callaway, 1993; Hovy, 1993; Reiter, 2000; Theune, 2000]:

- автоматическое порождение, основанное на использовании готовых текстовых фрагментов (шаблонов), т.е. **шаблонных технологий** (*templates technologies*),
- автоматическое порождение, основанное на использовании лингвистических знаний, т.е. **лингвистически мотивированных технологий** (*linguistically motivated technologies*).

Системы, использующие шаблонные технологии, работают с ЕЯ информацией как со строкой символов. Они могут создавать лишь стереотипные тексты из заранее подготовленных текстовых фрагментов (предложений), не позволяют менять формат и содержание порождаемого текста и имеют ряд других ограничений. Это обусловило необходимость поиска более глубоких подходов к созданию систем автоматического порождения, которые не будут иметь недостатков, присущих шаблонным методам.

Системы, использующие лингвистически мотивированные технологии, предназначены для создания текстов имеющих относительно свободное содержание, которое не может быть задано в виде готовых текстовых фрагментов. Они основанные на использовании лингвистических правил, которые применяются для эксплицитного описания знаний о структуре содержания и об устройстве порождаемого текста, а также знаний, которые позволяют выразить это содержание языковыми средствами.

2. Архитектура системы автоматического порождения англоязычных деловых электронных писем

Общая архитектура созданной системы порождения включает две части:

- ресурсы порождения, которые описывают знания о языке необходимые для синтеза текстов;
- обрабатывающий компонент, который реализуется как интерпретатор ресурсов порождения и организатор всего процесса работы системы.

В основу представленной системы были положены следующие основные принципы:

1) принцип модульности структуры системы;
2) принцип взаимодействия детерминированных и вероятностных правил в процессе порождения текста, который реализуется в основных модулях системы.

Многочисленные опыты в области моделирования процесса создания текстов с помощью ЭВМ показали, что систему автоматического порождения так же целесообразно организовать в виде набора модулей, в каждом из которых к обрабатываемой информации добавляются знания определенного лингвистического уровня. На начальном этапе применяются знания об организации логико-семантической структуры текста, затем об организации синтаксической структуры предложений. В конце процесса порождения решаются вопросы согласования лексических единиц. Таким образом, автоматическое порождение текста может быть определено как лингвистически мотивированный процесс построения текста на естественном языке последовательным преобразованием его порождаемой структуры от семантического уровня к текстовому.

В языке все уровни лингвистического описания (семантический, лексический и грамматический) тесно связаны друг с другом, явления одного уровня влияют на явления другого уровня и наоборот. Однако отношения такой сложности тяжело поддаются алгоритмизации, так как очень трудно вскрыть все зависимости и обеспечить их координацию в системе. Для облегчения задачи создания системы порождения текстов, процесс синтеза текста упрощается до последовательного применения модулей. Такой подход соответствует известной в теории информационных систем идеи «конвейера обработки данных».

Структурно-функциональная схема системы порождения включает три основных модуля: модуль планирования содержания, модуль языкового оформления и модуль редактирования переменных данных (рис. 1).

Модуль планирования содержания текста состоит из двух подмодулей, работающих в тесном взаимодействии. Первый подмодуль определяет, какая информация будет участвовать в

порождаемом тексте, а второй – порядок следования информации в документе. Результатом работы первого модуля является логико-семантическая модель текста делового письма.

В *модуле языкового оформления текста* происходит выбор лексических и грамматических средств, представленных в базе знаний системы, которые необходимо использовать, чтобы выразить на английском языке содержание, сформированное в первом модуле. На данном этапе порождения осуществляется выбор лексических единиц, а также их грамматических форм, выполняется морфологическое согласование между членами грамматических групп, а также проверка соответствия выбранных слов структурно-семантическим и лексико-семантическим правилам сочетаемости. Таким образом, данный модуль переводит логико-семантическое представление, построенное предыдущим модулем, в текст на английском языке.

В *третьем модуле* происходит окончательное оформление текста, подстановка в текст письма даты, а также заполнение количественных данных.

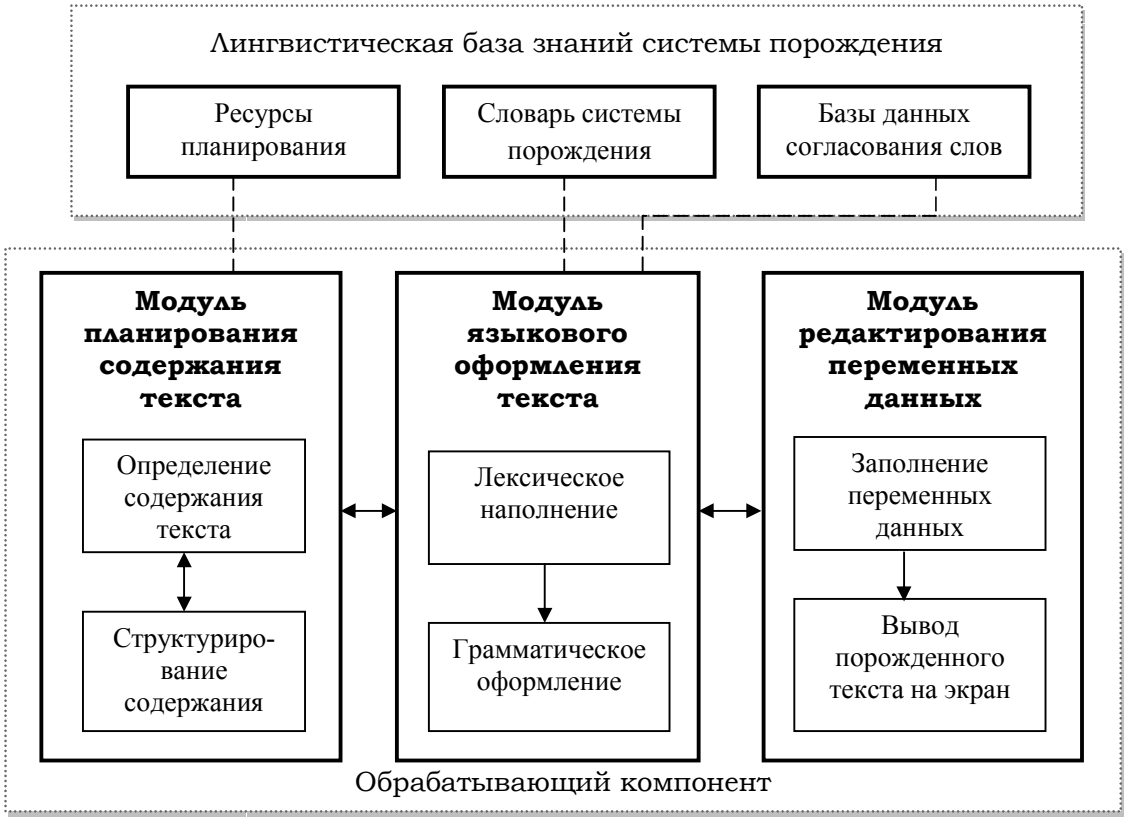

Рисунок 1 – Структурно-функциональная схема порождения

3. Лингвистическая база знаний системы автоматического порождения

Для практической реализации вышеописанных модулей на основе анализа предварительно составленного корпуса англоязычных деловых электронных писем была создана лингвистическая база знаний, которая включает:

- специализированный англоязычный словарь, содержащий свыше 6000 словарных статей, в которых подробно описываются семантические и морфологические характеристики каждой лексической единицы;
- логико-семантические формулы, которых определяют порядок формирования содержания текстов англоязычных деловых электронных писем, по выбранной пользователем теме;
- семантико-синтаксические формулы текстов заданной тематики, на основании которых происходит построение синтаксической структуры предложений, а также их лексическое наполнение с опорой на специализированный англоязычный словарь;

- базы данных, содержащие правила согласования лексических единиц, в которых зафиксированы характерные для официально-делового стиля нормы лексической сочетаемости.

4. Принципиальный алгоритм порождения англоязычных деловых электронных писем

На основе разработанной структурно-функциональной схемы вероятностно-алгоритмической модели была предложена следующая принципиальная схема алгоритма порождения деловых электронных писем, представленная на рис 2.

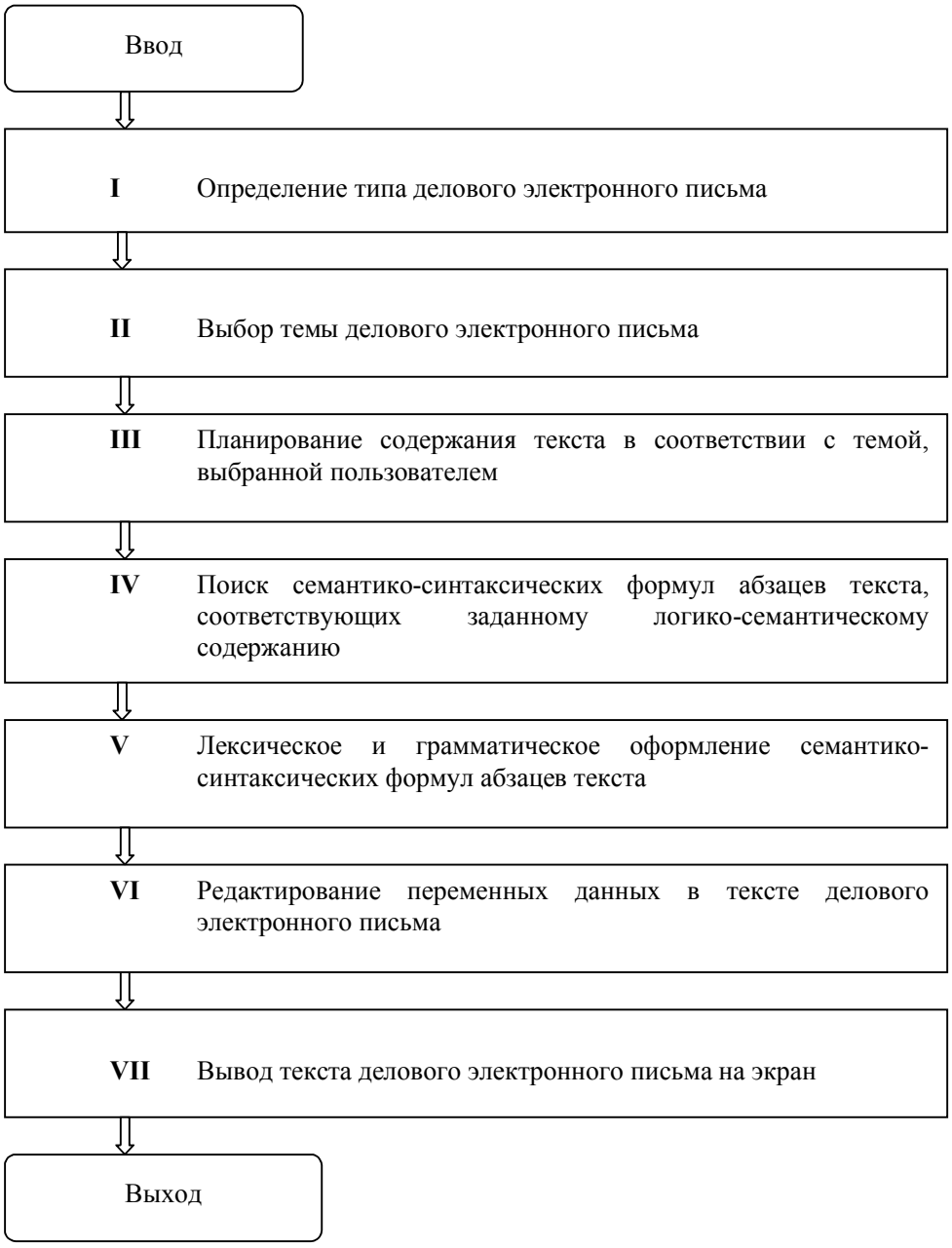

Рисунок 2 – Основные блоки принципиального алгоритма автоматического порождения текста

Отличительной чертой разработанного алгоритма является то, что в нем учитывается принцип взаимодействия детерминированных и вероятностных факторов в процессе порождения текста. На этапе планирования содержания осуществляется вероятностный выбор

одной из возможных логико-семантических формул текстов, таблиц основного статического содержания и семантико-синтаксических формул абзацев, на основании которых происходит планирование и структурирование содержания будущего текста (блоки III–IV). На этапе языковой реализации текста выбор соответствующих лексических единиц из базы знаний системы при заполнении семантико-синтаксических формул осуществляется вероятностным образом, а грамматическое оформление предложений текста происходит по строго детерминированным правилам (блок V).

5. Тестирование системы и оценка полученных результатов

Компьютерный эксперимент для верификации описанной выше системы порождения англоязычных деловых электронных писем был проведен с помощью программы на языке C #. Описываемая экспериментальная программа состоит из трех структурных частей:

- *Предварительная настройка*, которая отвечает за выбор языка пользовательского интерфейса. Созданная система порождения англоязычных деловых электронных писем поддерживает интерфейс на двух языках: английском и русском.
- *Основная часть* реализует модули “Планирование содержания” и “Языковое оформление” системы автоматического порождения.
- *Заключительная часть* реализует модуль “Редактирование переменных данных”, в котором происходит выбор даты, а также заполнение количественных данных.

Для создания текста англоязычного делового электронного письма с помощью разработанной системы порождения пользователю необходимо выполнить следующую последовательность действий:

- выбрать тип делового письма, например: рекламация, просьба, заказ или благодарственное письмо;
- выбрать тему будущего текста из множества тем, которые выводятся на экран компьютера;
- выбрать на панели управления системы опцию “*Создать*” и на экране появится текст англоязычного делового электронного письма, порожденный системой;
- заполнить необходимые переменные данные в диалоговом окне “*Настройки письма*”, которое автоматически появляется на экране компьютера.

Полученные в ходе тестирования результаты, свидетельствуют о том, что созданная система автоматического порождения является эффективной:

a) позволяет оперативно порождать тексты англоязычных деловых электронных писем по выбранной пользователем теме;
b) обеспечивает достаточно высокое качество создаваемых текстов и их соответствие международным стандартам составления деловых документов (пример делового электронного письма, порожденного системой, представлен на рис. 3);
c) позволяет осуществлять выбор языка пользовательского интерфейса (русский / английский) и может быть использована даже специалистами, не владеющими в совершенстве английским языком.

Дальнейшее усовершенствование разработанных лингвистических технологий автоматического порождения текстов возможно в нескольких направлениях. Прежде всего, это расширение границ их применения и использование данных технологий для создания не только англоязычных деловых документов, но и текстов, принадлежащих к другим функциональным стилям и жанрам. Полученные данные могут стать основой для создания многоязычных систем порождения, позволяющих синтезировать тексты на различных языках.

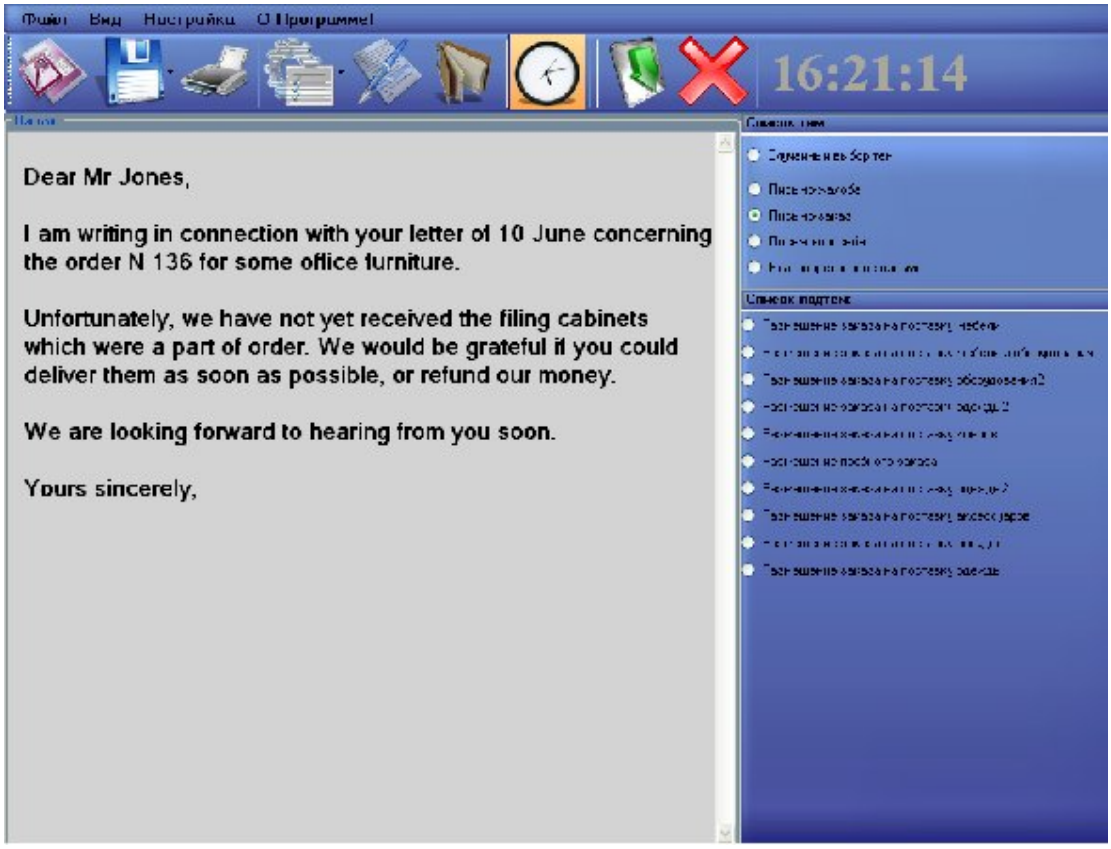

Рисунок 3 – Пример работы компьютерной программы автоматического порождения

Исследование выполнялось и по индивидуальному гранту, выделенному Министерством образования Республики Беларусь (номер госрегистрации 20090536 от 10.02.2009 г., сроки выполнения – 2009–2010 гг.), для изучения проблемы автоматического порождения англоязычных деловых документов с помощью компьютера.

Библиографический список

[Болдасов, 2003] Болдасов, М.В. Парадигмы генерации ЕЯ текстов в инструментальной среде DEMLING / М.В. Болдасов // Труды международной конференции Диалог 2003 по компьютерной лингвистике и интеллектуальным технологиям. – Протвино, 2003. – С. 66–75.

[Соколова, 2005] Соколова, Е.Г. Генерация текстов на естественном языке – состояние вопроса и прикладные системы / М.В. Болдасов, Е.Г. Соколова // НТИ. Сер. 2. Информационные процессы и системы. –2005. – № 10. – С. 12–22.

[Callaway, 1993] Callaway, C. Narrative Prose Generation: PhD thesis / C. Callaway // Department of Computer Science, North Carolina State University. – Raleigh, 2000. – 284 p.

[Hovy, 1993] Hovy, E. Automatic generation of formatted text / E. Hovy // Artificial Intelligence. – № 63. – 1993. – P. 341–385.

[Reiter, 2000] Reiter, E. Building Natural Language Generation Systems / E. Reiter, R. Dale. – Cambridge : Cambridge University Press, 2000 – 243 p.

[Theune, 2000] Theune, M. From date to speech: language generation in context : PhD thesis / M. Theune. – University of Eindhoven. The Netherlands, 2000. – 219 p.