Т. В. Бусел

СОВРЕМЕННЫЕ ПОДХОДЫ К ПОРОЖДЕНИЮ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Проблема порождения связного текста на естественном языке приобрела особую актуальность в связи с развитием перспективных информационных и коммуникационных технологий, в частности глобальной компьютерной сети Интернет, а также необходимостью быстрого ввода, обновления, обмена текстовой информацией в режиме реального времени. Решение данной проблемы возможно через моделирование и последующую автоматизацию процесса порождения текстов на естественном языке.

Несмотря на достаточно длительную историю развития подходов к порождению текстов, разнообразие предложенных методов, эта проблема еще не до конца решена. Большинство действующих компьютерных моделей используют шаблонные технологии, главная особенность которых заключается в том, что содержание будущего текста в них представлено в виде готовых фрагментов, что касается лингвистически мотивированных технологий, то они находятся в стадии исследования.

Моделирование процесса порождения текстов на естественном языке предполагает решение вопросов, связанных с изучением жанровой и коммуникативных характеристик текстов, риторических приемов организации

содержания текста, языковых средств выражения связаности текста, формализацией грамматики и лексических описаний. Таким образом, рассматриваемая проблема находится в неразрывной связи с рядом актуальных вопросов современного теоретического языкознания и компьютерной лингвистики: способами представления смысловой структуры текста и его единиц; закономерностями построения текста; вариативностью языкового выражения его содержания; определением качественного своеобразия функционирования языковых единиц различных уровней в тексте; построением типологии текстов.

В настоящее время известны многочисленные подходы к процедуре порождения текстов с помощью компьютера, которые подробно освещаются в работе американского ученого Эдварда Хови [1]. Зависят они в основном от того, для какой цели создается текст. По степени сложности и выразительности, существующие методы порождения сообщений принято подразделять на четыре класса:

1. Canned-based method. Для порождения сообщений создаются таблицы неизменяющихся шаблонов, которые используются системой в зависимости от ситуации. Этот метод предназначен для порождения простых цепочек слов (например, сообщение об ошибке в работе программного продукта т.д.).

2. Template-based method. Данный метод связан с созданием различного рода диалоговых систем, применяемых в справочных и обучающих системах, как правило, это шаблонные системы (template systems), которые используют готовые реплики или комбинируют готовые фрагменты текста таким образом, что они занимают заданные позиции в дискурсе или стереотипном тексте.

3. Phrase-based method. Более сложный метод, при которым используются универсальные фразовые шаблоны на синтаксическом уровне и на уровне дискурса, поэтому их также называют «планами текста» (text plans). В таких системах, фразы строятся в соответствии с определенной моделью (например, подлежащие + сказуемое + дополнение), а затем каждая составляющая данной модели находит свое воплощение в соответствии с грамматическими правилами, заложенными в систему. Процесс построения модели предложения завершается тогда, когда каждая его составляющая выражается конкретным словом или сочетанием слов. Такие системы являются достаточно эффективными, но имеют определенные ограничения, вызванные необходимостью детально описывать межфразовые связи и способы их реализации, для построения грамматически правильных предложений.

4. Feature-based method. Это наиболее сложный метод, он требует привлечения обширных лингвистических знаний, но, в то же время, он и наиболее привлекателен. Синтез сообщения осуществляется на основе набора свойств (грамматических признаков). При таком подходе предложение определяется набором характеристик составляющих его слов (например, наличие/отсутствие отрицания, настоящее/прошедшее время) и правилами их сочетаемости.

Архитектура современных систем генерации текстов на естественном языке (ГЕЯ), как правило, представлена тремя основными составялющими: *оболочка, планировщик и лингвистический компонент*. Рассмотрим их основные функции, опираясь на работы [2, с. 12–14; 3, p. 53–74].

О б о л о ч к а (underlying application program) определяет назначение СТГ и характер баз знаний, из которых черпается информация для построения текста. Оболочка выполняет две основные функции: инициирует процесс генерации и определяет цели, которые должны быть достигнуты высказываниями.

П л а н и р о в щ и к определяет пути достижения высказываниями поставленных оболочкой целей в данном предметном контексте. Он обеспечивает:

1) выбор информации, которая должна быть выражена или опущена;

2) определение того, как она должна быть представлена (как событие, например, “the economy developed” или как объект, например, “the development of the economy” и т.д.);

3) выбор способа взаимодействия с лингвистическими данными (лексика открытых классов, синтаксические конструкции). В частности, планировщик выполняет следующие задачи:

– структурирование текста – определение порядка пропозиции и границ предложений в выходном тексте;

– выбор лексики (без ограничений на синтаксический класс);

– построение синтаксической структуры предложений выходного языка;

– языковое оформление отношений кореференции (анафора, дейксис, эллипсис).

Каждая из перечисленных выше задач планировщика только теоретически может рассматриваться полностью изолированно от других. При экспериментальном моделировании, особенно в составе СТГ, их функции перемежаются.

Л и н г в и с т и ч е с к и й к о м п о н е н т порождает тексты в соответствии со спецификациями планировщика. Он обеспечивает грамматическую правильность текста и принимает большую часть, если не все синтаксические и морфологические решения. Полностью обеспечивает процессы синтаксического и морфологического синтеза текста на основе синтаксической структуры.

Процесс ГЕЯ принято представлять с использованием известной в теории информационных систем идеи конвейера обработки данных. Путем обобщения опыта создания действующих систем ГЕЯ, построена схема, представленная на рис. 1, которая отражает общую картину преобразований в системе ГЕЯ.

Из схемы видно, что система генерации состоит из трех относительно независимых блоков. Макропланирование – построение плана текста. Микропланирование – построение планов предложений и языковое оформление – реализация построенных планов предложений средствами конкретных ЕЯ. Рассмотрим более подробно каждый из этапов генерации.

Основная цель этапа м а к р о п л а н и р о в а н и я – создание плана текста. Для этого из входных данных система выбирает данные, релевантные поставленной коммуникативной цели, и организует их в последовательность изложения в будущем тексте.

Определение вида входных данных является кардинальным вопросом для ЛМ систем. Существуют три вида возможных входов для систем ГЕЯ – числовые данные, структурированные объекты и логические формулы. Более конструктивным является изучение входов, с которыми работают экспериментальные ЛМ системы. Можно выделить три вида таких входов.

1. Базы данных. 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», т.е. такой системы, которая позволяет построить семантическое представление из предлагаемых интерфейсом понятий на основе внутренней речи человека. Этот процесс называется “symbolic authoring” [4, с. 570].

3. Представление знаний на формальном языке.

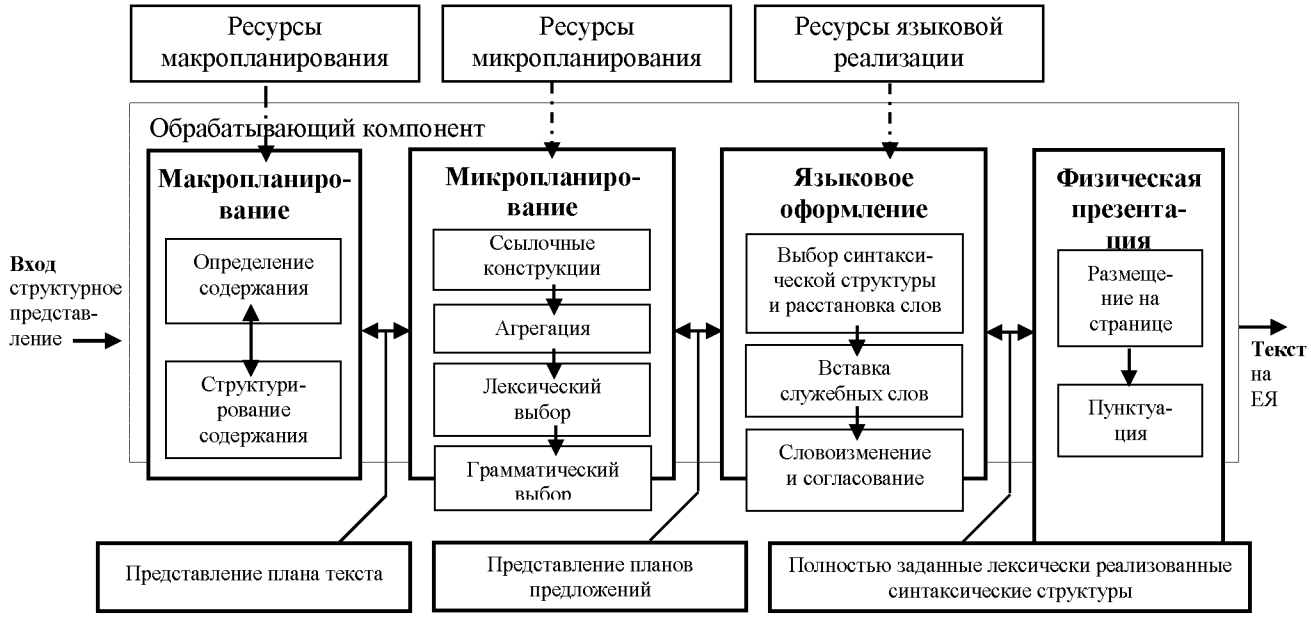

Рис. 1. Общая схема генерации

План текста – это представление информации, составляющей содержание будущего текста, организованное в виде единой структуры, которая может быть отображена в структуру текста на ЕЯ. Для представления этой информации может использоваться концептуальное представление, состоящее из объектов и отношений между ними. Объекты концептуального представления – это экземпляры сущностей ПО, порожденные из них согласно информации, представленной системе генерации, отношения между объектами – это отношения между соответствующими сущностями модели ПО, которые планируется рассмотреть в создаваемом тексте.

Способ записи концептуального представления определяется решаемой задачей каждой конкретной системы генерации. Отношения концептуального представления имеют предикативную природу, т.е. каждому отношению можно сопоставить предикат, и представить в виде фрейма, в котором каждый из объектов отношения имеет фиксированную роль. Поэтому отношения легко интерпретировать как планы отдельных высказываний – сообщения.

Сообщения являются концептуальными планами отдельных высказывание текста. Они могут представляться различными средствами с разной степенью общности, начиная с представления сообщений как идентификаторов обсуждаемых в тексте тем, и кончая их представлением в виде логических формул и шаблонов высказывание.

Концептуальное представление не дает исчерпывающего ответа о порядке изложения содержащейся в нем информации, о том какие фреймы должны быть выделены из этого представления, и в каком порядке они должны быть включены в результатирующий текст. Поэтому оно не может считаться планом текста.

Вопрос о том, по каким правилам строится текст, как выстроить отдельные высказывания текста, чтобы достичь желаемого воздействия на адресата, изменения его знаний, убеждений и намерений, изучает риторика. Поэтому вопрос построения плана текста – это вопрос построения его рито-рической структуры.

Риторическую структуру имеет текст, призванный воздействовать на адресата. Не случайно в античности риторика обслуживала речи ораторов. Но воздействие на адресата – не единственный мотив организации текста. Стратегия построения текста может также определяться самими данными, однако все они могут быть представлены в терминах риторической структуры, которая в данном случае рассматривается скорее как форма представления структуры текста.

Основные принципы риторического моделирования структуры текста сформулированы в теории риторических структур – *Rhetorical Structure Theory (RST)*. Как отмечают В. Манн и С. Томпсон «риторическая структура – это древесное представление определяющее, как будет организован создаваемый текст» [5, p. 243–281]. Его терминальные вершины – сообщения соответствуют отдельным высказываниям текста. Внутренние вершины риторического представления описывают, как сообщения структурированы вместе и связаны друг с другом. Внутренние вершины задают риторические отношения между связанными с ними фрагментами текста, например, цель, причина, последовательность, уточнение, побуждение, разрешение и т.д. Каждая внутренняя вершина разделяет свое содержание как минимум на две части: главную (nucleus) и второстепенную (satellite).

Группировка текстовых фрагментов риторическими отношениями накладывает ограничения на будущее разделение порождаемого текста на абзацы и предложения. Существование единого покрытия текста риторической структурой соответствует ощущению человеком его цельности.

Существуют два основных подхода к генерации риторической структуры текста: «основанный на планирующих операторах на предикативных схемах» [6, с. 5–6].

Первый подход более теоретически мотивирован – задача выбора структуры текста полностью перекладывается на ресурсы генератора – планирующие операторы.

Второй – более подходит для практических задач. Он задает предварительные шаблоны структуры текста – предикативные схемы, уточняемые при построении структуры конкретного текста. Данный подход, впервые предложенный К. Маккьюин [7, с. 2–42], основан на замечании, что для конкретной коммуникативной цели люди регулярно используют одни и те же виды информации в одном и том же порядке. Предикативная схема – это шаблон, который определяет, как должен быть организован текст, используя для этого более мелкие схемы, сообщения и дискурсные отношения между ними.

После построения плана текста и сообщений выполняются задачи микропланирования. Целью данного этапа является составление плана отдельных предложений генерируемого текста на основе сообщений с учетом общей структуры текста.

Для представления планов предложений на практике используются различные техники:

1. План предложения может состоять из набора уже готовых фрагментов высказывания, которые на этапе языкового оформления нужно лишь немного доработать для лучшей согласованности частей предложения.

2. План предложения может представлять собой полуграмматическую структуру, содержащую отдельные фрагменты высказывания, грамматические единицы и семантические признаки.

3. План предложения может быть представлен семантической структурой. Своеобразие семантических представлений состоит в сочетании достаточно высокой абстракции и одновременно близости к ЕЯ, что позволяет рассматривать их как универсальное предъязыковое представление данных или стандарт представления планов предложений.

Семантическое представление предложения строится из одного или нескольких соседних сообщений с учетом окружающей их риторической структуры. Для того чтобы провести такое преобразование на этапе микропланирования, выполняются четыре основные задачи:

1. Агрегация, в ходе которой происходит объединение простых фраз в более сложные структуры предложений (простое сочинение, синтаксическое подчинение и т.д.).

2. Лексикализация концептов сообщения, т.е. выбор подходящих слов для выражения их содержания.

3. Вставка ссылочных конструкций. Для обеспечения лучшей слитности текста при повторном упоминании объекта в высказываниях для его идентификации выбираются различные слова или словосочетания (местоимения, дефинитные описания и т.д.).

Таким образом, на этапе микропланирования построенные сообщения преобразуются, учитывая их расположение в плане текста, в планы отдельных предложений.

На этапе языкового оформления эти планы реализуются средствами лексики и грамматики конкретного ЕЯ в грамматические структуры, которые затем преобразуются в предложения ЕЯ текста. Готовые предложения собираются в результатирующий текст.

Этот этап называется также поверхностной реализацией и базируется на положениях трансформационной грамматики, разработанной Н. Хомским, который разделил лингвистические представления на глубинные и поверхностные. Глубинное грамматическое (семантическое) представление фактически содержит план поверхностной грамматической структуры высказывания. На этапе поверхностной реализации генератор выбирает грамматические конструкции – функциональные роли (подлежащее, прямое дополнение и т.д.), определяет линейный порядок (определяемый грамматикой), части речи (существительное, глагол и т.п.), сложность предложения (простое, сложное) и окончательную форму слов (морфология), вставляет служебные слова (союзы, предлоги, артикли). Ресурсы этого уровня описывают лексический, морфологический и синтаксический уровни лингвистической модели языка.

Таким образом, в процессе генерации входное представление последовательно преобразуется между следующими лингвистическими уровнями: концептуальный уровень, семантический уровень, риторический уровень, синтаксический уровень и текстовый уровень. Считается, что первые три уровня описывают надъязыковые явления. Последние два уровня описывают явления, специфичные для конкретного языка. Генерация в такой уровневой модели может быть определена как лингвистически мотивированный процесс построения текста на ЕЯ последовательным преобразованием его порождаемой структуры от концептуального уровня к текстовому.

Анализ приведенных выше подходов к созданию систем генерации показывает, что данные модели, как правило, не рассчитаны на порождение текстов по заданному содержанию.

Нами была разработана вероятностно-алгоритмическая модель порождения текста, позволяющая порождать тексты деловых электронных писем различных типов на английском языке по заданному содержанию. Созданная модель является алгоритмической, относящейся к классу воспроизводящих лингвистических моделей, которые с одной стороны, имитируют структуру и поведение реальных лингвистических объектов, а с другой – позволяют воспроизводить эти объекты. Учитывая многоаспектность, многоуровневость порождаемых объектов, при создании данной модели мы стремились наиболее полно и точно отразить в ней реальные свойства языка.

Процесс генерации текста компьютерной системой происходит в две стадии. Первая определяет содержание и структуру будущего текста, это *стратегический компонент*, его еще называют “планировщиком” текста. Вторая стадия – *лингвистический* (или *тактический*) *компонент*, – определяет, как строить текст делового письма, какие лексические, синтаксические и коммуникативные средства естественного языка нужны для порождения текста.

На основе анализа различных текстов деловых писем была создана база данных для работы программы порождения. Для каждого текста делового письма в такой базе данных указаны:

1. Логико-семантическая формула текста, которая представляет собой линейную последовательность абзацев с определенным предметно-логическим содержанием.

2. Таблица основного статического содержания (ТОС), в которой указанны главные и второстепенные опорные слова, которые отражают главные субъекты и объекты ситуации, описываемой в тексте.

3. Алфавитно-частотный словарь текста.

4. Семантико-синтаксические формулы абзацев (СЕСФА) текста, представленные на специальном семантико-синтаксическом языке, в основе которого лежат семантические функции.

Подбор СЕСФА для заданного основного содержания и порядок их следования в пределах семантико-синтаксической формулы текста определяется двумя типами факторов:

а) вероятностными, выявленными в процессе статистического анализа следования абзацев с разным предметно-логическим содержанием в текстах деловых писем;

б) детерминированными, полученными в результате качественного изучения ТОС и СЕСФА.

Определяющими при этом являются факторы детерминированные.

При внедрении в промышленные системы обработки связных текстов созданная модель вероятностно-алгоритмического порождения текстов на естественном языке позволит граммотно и быстро порождать различные типы деловых документов по заданному содержанию, соответствующие международным стандартам, и может быть использована даже специалистами, не владеющими в совершенстве английским языком. Поскольку английский язык является основным средством международного делового общения во всем мире, данная вероятностно-алгоритмическая модель порождения может иметь широкое практическое применение.

ЛИТЕРАТУРА

1. *Hovy, E.* Language Generation / E. Hovy// Survey of the state of art in human language technology [Электронный ресурс]. – Режим доступа : http://www.isi.edu/natural-language/people/hovy/publications.html. – Дата доступа : 10.06.2008.

2. *Соколова, Е. Г.* Лингвистические компоненты в экспериментальных системах генерации текстов (по работам ученых США и Канады) / Е. Г. Соколова // НТИ. Сер. 2. Информационные процессы и системы. – 1993. – № 4. – С. 10–14.

3. *Bateman, J. A.* An overview of computational text generation / C. Butler, editor // Computers and Texts : An Applied Prospective. – Basil Blackwell, Oxford, England, 1992. – P. 53–74.

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

5. *Mann, W.* Rhetorical structure theory toward a functional theory of text organization / W. Mann, S. Thompson // Text. – 1988. – Vol.8. – № 3. – P. 243–281.

6. *Болдасов, М. В.* Генерация текстов на естественном языке – теории, методы, технологии / М. В. Болдасов, Е. Г. Соколова // НТИ. Сер. 2. Информационные процессы и системы. – 2006. – № 7. – С. 1–14.

7. *McKeown, K. R.* Text Generation: Using discourse strategies and focus constrains to generate natural language text / K. R. McKeown. – Cambridge University Press., 1986. – 246 p.