ПРОБЛЕМЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

Т. В. Бусел

РАЗРАБОТКА АВТОМАТИЗИРОВАННОГО МЕТОДА ПОРОЖДЕНИЯ ДЕЛОВЫХ ДОКУМЕНТОВ НА ОСНОВЕ ЛИНГВИСТИЧЕСКИХ ПРАВИЛ

В статье рассматриваются теоретические и практические вопросы, связанные с созданием систем автоматического порождения текстов на естественном языке, описывается общая организация системы: способы представления данных, теории и технологии реализации процесса порождения. Предлагается вероятностно-алгоритмический метод генерации деловых документов, который учитывает правила их логико-семантической, синтаксической и лексической организации, а также вероятностные факторы, влияющие на способы тематического развертывания и вербализации текстовых компонентов.

Появление и широкое распространение современных информационных и коммуникационных технологий значительно ускорило развитие научной области, известной как компьютерная лингвистика. Сфера приложений компьютерной лингвистики постоянно расширяется, появляются все новые фундаментальные задачи, связанные с моделированием разнообразных видов речемыслительной деятельности людей. Генерация текста на естественном языке является одной из наиболее сложных задач искусственного интеллекта и компьютерной лингвистики.

Лингвистические проблемы, возникающие при создании систем порождения текстов (СПТ), напрямую связаны со свойствами языка и обусловлены целым рядом факторов. В о - п е р в ы х, для естественного языка (ЕЯ) характерны непостоянность и неоднородность правил его описания. Большое разнообразие грамматических форм и правил часто не позволяет выделить некую единую структуру, необходимую для построения компьютерной системы. Более того, набор лингвистических правил, разработанных для языка с жестким порядком слов в предложении (как, например, английский язык), не может быть использован для флективно богатого языка (языки славянской группы).

В о - в т о р ы х, возникает проблема описания смысловой составляющей текста. Порождение текста требует создания мощного компонента искусственного интеллекта, который мог бы работать с лингвистическими базами знаний, адаптировать систему для новых знаний и обучаться.

При моделировании процесса порождения текста компьютером необходимо учитывать, что ЕЯ – это не просто набор слов, связанный грамматическими правилами. Приоритетной задачей является получение именно осмысленного текста, что, в свою очередь, приводит многих разработчиков к необходимости учета семантических связей не только между отдельными словами, но и между предложениями и даже документами. С этой точки зрения наиболее сложной и интересной является именно генерация текстов, реализация которой будет наиболее полно учитывать все важные смысловые связи в документе.

В настоящее время известны многочисленные подходы к процедуре порождения текстов с помощью компьютера, которые подробно освещаются в работе американского ученого Э. 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 и 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 котором 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 составляющими: *оболочка*, *планировщик* и *лингвистический компонент*. Рассмотрим их основные функции, опираясь на работы [2, с. 12–14; 3].

О б о л о ч к а (underlying application program) определяет назначение СПТ 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:

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

2) выбор лексики;

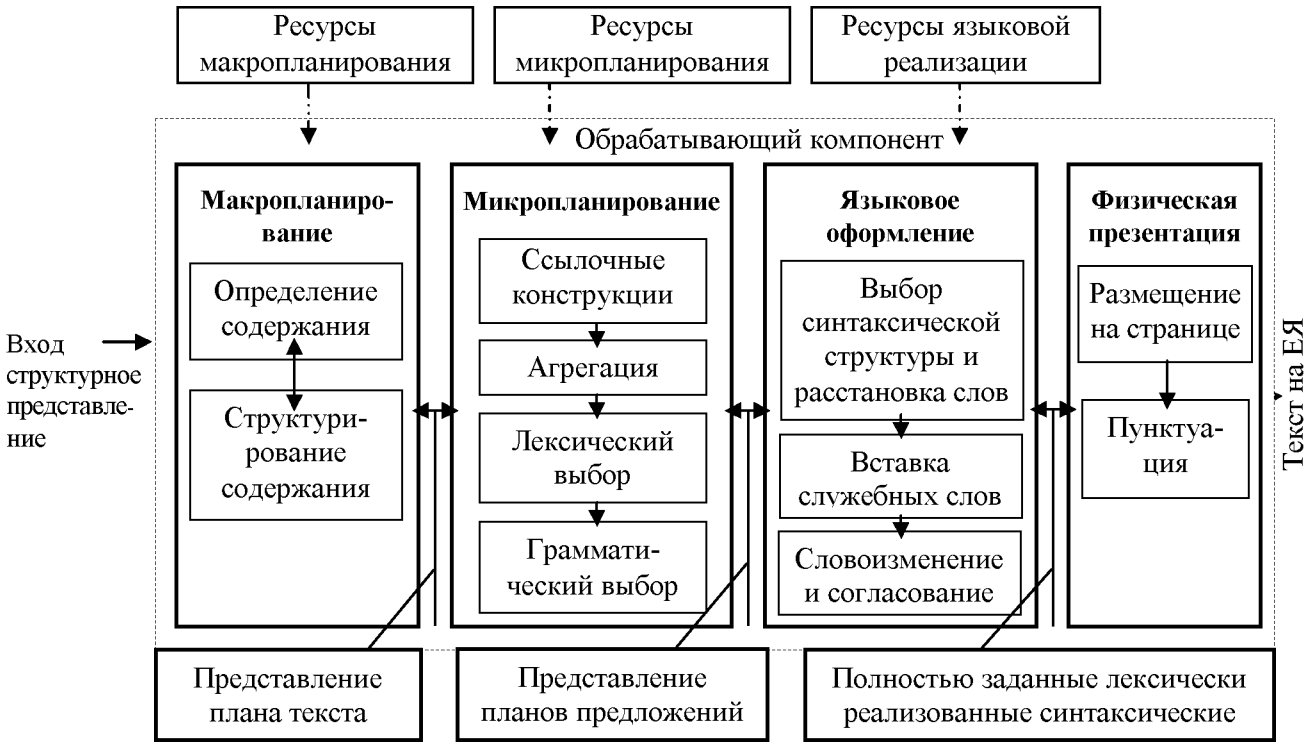
3) построение синтаксической структуры предложений выходного языка;

4) 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 СПТ, их функции перемежаются.

Лингвистический компонент 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 представленная на рисунке схема, которая отражает общую картину преобразований в системе ГЕЯ.



Общая схема генерации

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 макропланирования – 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. Как правило используются:

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» [цит. по: 4];

3) представление знаний на формальном языке.

План текста – это представление информации, составляющей содержание будущего текста, организованное в виде единой структуры. 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.

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. 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 три 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, с учетом их расположения в плане текста, преобразуются 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. 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, части речи (существительное, глагол и т.д.), 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, а последние два уровня – 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 методы, как правило, не рассчитаны на порождение текстов по заданному содержанию.

Нами был разработан вероятностно-алгоритмический метод порождения текста, позволяющий порождать тексты деловых документов различных типов на английском языке по заданному содержанию. 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 документа, какие лексические, синтаксические и коммуникативные средства естественного языка нужны для порождения текста.

На основе анализа различных текстов деловых документов были созданы алгоритм и база данных для работы программы порождения. Для каждого текста делового документа 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 приведены 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 разработанный вероятностно-алгоритмический метод порождения текстов на ЕЯ позволит грамотно и быстро порождать различные типы деловых документов по заданному содержанию, соответствующие международным стандартам. Поскольку английский язык является основным средством международного делового общения во всем мире, система, разработанная на основе предложенного вероятностно-алгоритмического метода, может иметь широкое практическое применение.

ЛИТЕРАТУРА

1. *Hovy, E.* Language Generation [Электронный ресурс] / E. Hovy // Survey of the state of art in human language technology. – 2017. – Режим доступа : http://www.isi.edu/natural-language/people/hovy/publications.html. – Дата доступа : 10.06.2017.

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 / J. A. Bateman // Computers and Texts : An Applied Prospective / C. Butler (ed.). – 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.

The article is devoted to theoretical and practical aspects of natural-language generation. It presents a new probably-algorithmic method, which takes into account determinative and probabilistic factors that play an important role in the text generation process and thus provide a more effective solution of the task.

Поступила в редакцию 02.10.18