М.С. Богомольская (Минск, МГЛУ)

ПОДХОДЫ К ПОРОЖДЕНИЮ ПРОЗАИЧЕСКОГО ТЕКСТА КОМПЬЮТЕРОМ

Проблема порождения связного текста очень сложна, поскольку текст является результатом Звзаимодействия большого числа факторов — лингвистических, психологических, логических, физиологических. В целом, технологии порождения прозаических текстов компьютером можно разделить на две большие группы: шаблонные технологии и лингвистически мотивированные технологии [1]. **Шаблонные технологии** относительно просты, надежны и находят широкое промышленное применение. Самые простые шаблонные системы используют готовые фрагменты написанного человеком текста без их дополнительной обработки. Более сложные системы позволяют задавать отдельные грамматические компоненты текста или комбинировать шаблонные высказывания и, таким образом, получать связный текст, используя при этом определенные лексические и грамматические знания о языке. Порожденные тексты выглядят вполне естественно, т.к. представляют собой последовательность фрагментов готового текста. Однако необходимо отметить, что подобные системы работают только с очень жесткими типами текстов. Использование шаблонов при порождении текста компьютером эффективно и целесообразно тогда, когда заранее известна структура порождаемого текста и существует возможность заранее определить его лексическое наполнение.

Системы, созданные на основе **лингвистически мотивированных технологий** ценны тем, что позволяют генерировать тексты с относительно свободным содержанием, которое не может быть задано при помощи готовых фрагментов текста. В данном случае источником содержания порождаемого текста являются данные, представленные в виде баз данных, баз знаний или выражений на формализованных языках. Тип входных данных не всегда предсказывает тип выходного текста, поэтому последний определяется извне пользователем. Подобные системы создают тексты одного типа, но в разных предметных областях, или в одной предметной области, но на разных языках.

Современные системы порождения текста состоят из следующих основных компонентов: оболочки, планировщика и лингвистического реализатора. **Оболочка** определяет назначение системы генерации и характер базы знаний, на основе которой происходит построение текста. Она выполняет две основные функции: инициирует процесс порождения и обуславливает цели, которые должны быть достигнуты в результате синтеза высказывания. **Планировщик** определяет пути достижения поставленных целей в данном предметном контек-сте. Он обеспечивает:

1) выбор информации, которая должна быть представлена в тексте;

2) определение того, как должна быть представлена эта информация;

3) выбор способа взаимодействия с лингвистическими данными. В частности, планировщик проводит структурирование текста, построение синтаксической структуры предложений и выбор соответствующей лексики. Опираясь на концептуальное представление текста, выработанное планировщиком, **лингвистический реализатор** осуществляет окончательный контроль за процессом порождения текста. Он принимает все окончательные морфологические и синтаксические решения и отвечает за грамматически правильное оформление текста.

В процессе генерации текста компьютером ученые выделяют три относительно независимых этапа:

1) макропланирование (построение плана текста);

2) микропланирование (построение плана каждого предложения текста);

3) языковое оформление предложений средствами конкретного языка [1; 2].

Рассмотрим эти этапы более детально. На этапе **макропланирования** компьютер принимает решение, какая именно информация из входных данных попадет в текст и как она будет организована. В данном случае он работает исключительно с предметными знаниями и общими способами организации содержания в тексте. Результатом этого этапа является план текста в терминах последовательности событий, например, запрос информации, предоставление информации или риторических решений. Определение вида входных данных является самым важным вопросом для лингвистически мотивированных систем. Обычно они используют представления данных, порожденные другими компьютерными системами для некоторых практических целей, а не созданные вручную разработчиками. Выделяют три вида входных данных [2, с. 175]:

1) **база данных** — особенность этого типа источника информации заключается в том, что данные не организованы для их передачи адресату. Тип текста, который можно построить на основе такой информации, и его структура, должны быть определены извне пользователем;

2) **семантическое представление**, то есть представление содержания текста, созданное с помощью системы интерфейсного типа «человек-компьютер»;

3) представление знаний на **формальном языке** (например, SQL или логических языках).

План текста — это представление информации, составляющей содержание будущего текста, организованное в виде единой структуры. С этой целью может использоваться **концептуальное представление,** состоящее из объектов конкретной области и отношений между ними. Способ записи концептуального представления определяется каждой конкретной системой генерации. Отношения концептуального представления имеют предикативную природу, то есть каждое отношение можно представить в виде фрейма, в котором каждый из объектов играет фиксированную роль. Поэтому отношения легко интерпретировать как сообщения, являющиеся концептуальными планами отдельных

высказываний текста. Они могут быть выражены различными средствами с разной степенью общности: от представления сообщений как идентификаторов обсуждаемых в тексте тем, до их записи в виде логических формул и шаблонов высказывания.

Стратегия построения текста может определяться и самими данными, которые могут быть представлены в терминах **риторической структуры.** Существует два основных подхода к генерации риторической структуры текста: 1) подход, основанный на планирующих операторах и 2) подход, базирующийся на предикативных схемах [3, с. 5–6]. Первый подход мотивирован теоретически, так как задача выбора структуры текста полностью перекладывается на ресурсы генератора (планирующие операторы). Второй подход больше подходит для решения практических задач.

После построения плана текста выполняются задачи этапа **микропланирования**. Его целью является составление плана отдельных предложений генерируемого текста с учетом общей структуры текста.

План предложения на практике может представлять собой:

1) набор уже готовых фрагментов высказывания, которые на этапе языкового оформления нужно лишь немного доработать для лучшей согласованности частей предложения;

2) полуграмматическую структуру, содержащую отдельные фрагменты высказывания, грамматические единицы и синтаксические признаки;

3) семантическую структуру [2, с. 177].

Семантическое представление предложения строится на основе 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.

Таким образом, на этапе микропланирования построенные сообщения модифицируются (с учетом их расположения в плане текста) в планы отдельных предложений.

На этапе **языкового оформления** эти планы преобразуются средствами лексики и грамматики конкретного языка в грамматические структуры, которые затем трансформируются в предложения текста. Таким образом, в процессе порождения прозаического текста его входное представление проходит последовательное преобразование на следующих лингвистических уровнях: концептуальном, семантическом, риторическом, синтаксическом и текстовом. Считается, что первые три уровня описывают общие для всех языков явления. Последние два уровня описывают явления, специфичные для конкретного языка. Генерация текста в такой многоуровневой модели может

быть определена как лингвистически мотивированный процесс построения текста на естественном языке в результате последовательного преобразования его порождаемой структуры от концептуального уровня к текстовому.

ЛИТЕРАТУРА

1. Соколова, Е.Г. Автоматическая генерация текстов на ЕЯ (портрет направления) / Портал по компьютерной лингвистике «Диалог» [электронный ресурс]. — 2009. — Режим доступа: http://www.dialog-21.ru/archive/2004/sokolova.htm. — Дата доступа: 15.04.2010.

2. Бусел, Т.В. О лингвистически мотивированных подходах к проблеме генерации текстов на естественном языке / Вестн. МГЛУ. Сер.1, Филология. — 2009. № 2. — С. 170–178.

3. Болдасов, М.В. Генерация текстов на естественном языке — теории, методы, технологии / НТИ. Сер. 2, Информационные процессы и системы. — 2006. — № 7. — С. 12–22.