

ПОЭЗИЯ И КОМПЬЮТЕР

Проблемами порождения текстов с помощью компьютеров исследователи занимаются с начала 70-х гг. За прошедшие годы по этим проблемам написано большое число книг, сборников и статей в различных журналах. Вопросам генерации текстов уделяют много внимания известные журналы. Например, «Artificial Intelligence Review», «Computational Linguistics», «International Journal of Human – Computer Studies», «Information Processing in Mind and Machine» и др. Ежегодно в разных странах мира по этим проблемам проводятся крупные конференции. Известнейшие университеты мира занимаются исследованиями по порождению текстов. Среди них Эдинбургский университет в Шотландии, Массачусетский, Стэнфордский, Колумбийский университеты в США, Монреальский университет и университет в Торонто (Канада) и много других. Тем не менее, до сих пор нет единой теории текстообразования, и эта теория находится в стадии становления [2; 3].

Известны многочисленные подходы к процедуре порождения текстов. Зависят они в основном от того, для какой цели создается текст. Большая часть таких алгоритмов связана с созданием различного рода диалоговых систем, используемых в справочных и обучающих системах. Следующая группа разработок имеет целью проверку различных лингвистических и литературоведческих теорий порождения предложения, абзаца, текста. Значительная часть работ направлена на генерацию практических документов (метеосводок, статистических отчетов о занятости населения, инструкций по регулировке механизмов и т.д.) или же небольших газетных сообщений о текущих событиях.

В Минском государственном лингвистическом университете также ведутся работы по моделированию на компьютере процессов текстообразования. Созданы алгоритмы и программы порождения французских сказок, английских и французских рекламных текстов, пословиц и поговорок [8]. Заканчивается работа по порождению на компьютере англоязычных электронных деловых писем [1].

Делаются попытки порождения и стихотворных текстов. Известный французский поэт Поль Валери говорил, что стихотворение создается не из идей и не из чувств, оно создается из слов. Иными словами, методы, используемые для порождения прозаических текстов вполне могут быть использованы и для порождения поэтических текстов. Но в то же время филологи, изучающие поэтические тексты, отмечают до 13 различий между прозаическими и поэтическими текстами [9, с. 35–46]. На уровне плана выражения это такие различия:

1. Стихотворная речь делится на сопоставимые между собой единицы (стихи), а проза представляет собой сплошную речь.

2. Каждый стих обладает внутренней мерой (метром), а проза ею не обладает.

Среди отличий в плане содержания можно отметить следующие:

1. В стихотворном тексте преобладает образное выражение смыслов, а в прозаическом их прямое выражение.

2. Стихотворная речь содержит, в основном, образную (лирическую фабулу), а прозаический текст характерен предметной (фактологической) фабулой.

3. Стихотворная речь характерна неоднозначностью её интерпретации, а прозаическая речь воспринимается, как правило, однозначно.

Отвлекаясь от различий в плане выражения, рассмотрим как попытались моделировать процесс стихосложения на компьютере. Первые попытки такого текстопорождения были проведены в г. Мишшенхиллз (штат Калифорния). Энтузиаст Розмари Уэст создала компьютерную программу POETRY GENERATOR/генератор стихов [4]. Автор следующим образом описывает её работу. В компьютере находится некоторый богатый словарь, где каждое слово имеет набор грамматических признаков и набор простых фраз, которые могут быть объединены в строки. Например, пусть из множества фраз, хранящихся в памяти компьютера, выбрано следующее четверостишие:

Дерево погружает голые пальцы
В черное озеро льда
Словно три серых гуся
Сползают по снежному склону.

Автоматически в каждой строке выделяются члены предложения: «дерево» – подлежащее, «погружает» – сказуемое, «голые пальцы» – дополнения и т.д. После этого на выделенные позиции подставляются слова таких же классов слов из большого словаря системы порождения. При этом выбор слов осуществляется компьютером случайным образом. В результате может получиться, например, другое четверостишие такой же структуры:

Женщина прячет пять серых котят
Под старым ржавым рыдваном
И тут же печальные клоуны
Входят в ваш музей страданий.

В качестве таких исходных четверостиший, закладываемых в компьютер, автор программы использовала части своих собственных стихов.

Другой алгоритм, созданный в Белфасте (штат Мэн) Томасом А. Истоном предусматривал симбиоз человека и компьютера. Была создана компьютерная программа THUNDER THOUGHT («удар мысли»), которая, как и вышесказанная, имела большой словарь, где каждое слово имело набор грамматической информации, и скелетные структуры предложений (строк). Компьютер, опять по датчику случайных чисел порождает заданное количество строк, а автор программы последовательно, на экране компьютера, их редактирует. Таким образом, было 110 стихотворений, из которых 32 были опубликованы. Ниже приводится одно из таких порожденных и отредактированных стихотворений.

Слаб от страданья необласканного «я»,
Любил он физику
И обнимая её каноны
Охватывал руками купола из линий
Антенн гигантских,
И с ревом на орбиту он вознесся,
И возвестил приход его сигнал, кричащий
Миллиметровыми волнами
По пространству.

Более серьезный и лингвистически более обоснованный подход компьютерного порождения стихотворений был продемонстрирован в России в 1975–1976 гг. Б.Г. Кацем [10]. Как и в предыдущих работах по порождению стихотворений, основой алгоритма явился словарь, состоящий из нескольких сотен слов, взятых из сборника стихотворений О.Э. Мандельштама «Камень». Каждое слово этого словаря имело грамматическую информацию о классе слова, числе, роде, времени, а также данные о метре и рифме. Метрическая информация задавалась двумя числами: 1) количеством слогов в слове до ударного слога; 2) число слогов после ударного слога. Словарь состоял из IV частей. В часть I были включены существительные и местоимения, которые смогут выступать в качестве подлежащего будущих компьютерных строк. Все прилагательные и притяжательные местоимения входили во II часть словаря. Они должны были выступать в роли определений. Часть III включала глаголы. Они должны выступать в качестве сказуемых будущих строк. К части IV были отнесены наречия, существительные с предлогами. Предполагалось, что на их основе будут формироваться обстоятельства (места, времени, образа действия).

В процессе порождения предполагалось, что строка стихотворения должна состоять из одного подлежащего, одного сказуемого, нескольких определений и обстоятельств. Некоторые из этих составляющих в процессе порождения могут быть опущены. Перед заданием для порождения компьютеру задавались данные о числе строк в строфе, какие строки рифмуются. Порождение строки начинается с права – налево. Методом случайного выбора из словаря подбираются последние рифмующиеся слова строк. Другие слова каждой строки также подбираются случайным образом с соблюдением определенных грамматических правил, первоначально заданных в порождаемой строке.

По данному алгоритму было порождено более 20 стихотворений на русском языке. Ниже приводится одно из них.

Вновь в кустах горят ресницы
Вечер хрупкий светлый злой
На столе желтели птицы
Взор играет за рекой
Лодка далеко краснеет
На закате соловьи
Вновь высокие белеют
Стены вечером твои.

Как видно, хотя смысла единого здесь нет, зато здесь есть рифма.

Многие исследователи отмечают, что при порождении текстов на естественном языке необходимо более серьезно относиться к целому ряду лингвистических особенностей порождаемых текстов. И, в частности, отмечается далее, что любая попытка абстрагироваться от индивидуальной системы порождения текста на естественном языке в пользу более обобщенной архитектурной специфики сталкивается на своем пути с серьезными препятствиями. И это действительно так.

Попытка учета индивидуальных речевых особенностей автора была осуществлена в Минском государственном лингвистическом университете при порождении французских текстов песен под Шарля Азнавура. Для построения словаря модели порождения были взяты 10 таких текстов песен. Словарь содержал все слова исследованных текстов в виде словоформ с указанием грамматических признаков (род, число, одушевленность, время, лицо, число и т.д.), семантического признака (принадлежность к определенному семантическому подклассу) и структурные признаки (количество слогов в слове, место ударного слога, номер рифмы). Для текста каждой песни были вычислены по статистической методике [6] главные опорные слова (ГОС) каждого текста и второстепенные опорные слова (ВОС) этих текстов. Они явились как бы заданием (темой) для порождения будущего текста.

Далее с опорой на один из методов описания логико-семантической структуры текста [5] были построены логико-семантические формулы каждой строки и строфы (4 строки) исследуемых текстов и полные логико-семантические формулы 10 исходных стихотворений. Затем с использованием специального семантико-синтаксического языка для записи текстов в памяти компьютера были созданы семантико-синтаксические формулы исследуемых текстов стихотворений.

Процесс порождения начинался с задания количества строф в будущем стихотворении и типа рифмы (aabb, abab и т.д.). После этого задавалась тема будущего стихотворения в виде случайного выбора одного из 10 ранее вычисленных наборов ГОС и ВОС. Также по датчику случайных чисел выбиралась одна из семантико-синтаксических формул будущего стихотворения. Ее заполнение осуществлялось из словаря системы порождения.

С помощью алгоритмического языка C++ была написана программа порождения стихотворений. Было порождено 5 французских стихотворений, основанных на лексике текстов песен Шарля Азнавура с соблюдением логико-семантической последовательности строф, типичных для исследованных текстов стихотворений.

Анализ порожденных текстов показал, что основные ошибки заключались в грамматическом согласовании рядом стоящих в строке слов. Из-за небольшой величины словаря и, тем самым, небольшого количества слов, входящих в лексико-семантические группы, нарушалось и семантическое согласование рядом стоящих слов.

Как было отмечено выше (стр. 2), одним из отличий поэтических текстов от прозаических является наличие в первых определенного эмоционального настроения. Сейчас на кафедре информатики и прикладной лингвистики МГЛУ идет работа по совершенствованию созданного алгоритма порождения стихотворного текста. Для этого каждая словоформа словаря снабжается и еще одним типом информации – принадлежность к определенному эмоциональному настрою («печальный», «радостный», «бодрый», «нежный», «агрессивный (злой)», «светлый», «тоскливый»).

Второе направление совершенствования связано с построением логико-семантической структуры создаваемого стихотворного текста. Для порождения не будет браться одна из готовых, существующих у поэта логико-семантических формул стихотворения (и соответственно – семантико-синтаксическая формула этого стихотворения). Такая формула будет состояться из последовательности строф, которые в стихах моделируемого поэта следуют наиболее часто друг за другом [7, с. 214–219]. Создаваемая модель порождения стихотворного текста будет проверяться на стихах Сергея Есенина.

Литература

1. Бусел Т.В. Вычислительные методы этапов генерации текстов // Материалы ежегодной научной конференции преподавателей и аспирантов университета, Минск 21–22 апр. 2009 г.: в пяти частях. Часть пятая. Минск, 2010. – С. 3–5.
2. Болдасов М.В., Соколова Е.Г. Генерация текстов на естественном языке – состояние вопроса и прикладные системы // НТИ. Серия 1. Информационные процессы и системы. – 2005. – № 10. – С. 12–23.
3. Болдасов М.В., Соколова Е.Г. Генерация текстов на естественном языке – теории, методы, технологии // НТИ. Серия 2. Информационные процессы и системы. – 2006. – № 7. – С. 1–15.
4. Дьюдни А.К. Компьютер пробует свои силы в прозе и поэзии // В мире науки, 1989. – № 8. – С. 88–92.
5. Зубов А.В. специфика русских текстов по употребительности в них абзацев с различным предметно-логическим содержанием // Квантитативная лингвистика и автоматический анализ текстов. Уч. записки Тартуского гос. ун-та. Тарту. Вып. 711. – С. 23–38.
6. Зубов А.В. Статический аспект содержания текста и его формальное представление // Квантитативная лингвистика и автоматический анализ текстов. Уч. записки Тартуского гос. университета. Тарту, 1986. – № 745. – С. 75–94.
7. Зубов А.В., Зубова И.И. Основы искусственного интеллекта для лингвистов: учеб. пособие. М., 2007.
8. Зубов А.В., Зубова И.И. Синергетические аспекты текстопорождения // Древняя и новая Романия: сб. ст., посв. памяти Р.Г. Пиотровского. СПб., 2010 (в печати).
9. Казарин Ю.В. Филологический анализ поэтического текста: учеб. для вузов. М., 2004.
10. Кац Б.Г. О программе, сочиняющей стихи // Автоматика и телемеханика. М., 1978. – № 2. – С. 151–156.