

ФОРМАЛИЗАЦИЯ ПРОЦЕССА ПОРОЖДЕНИЯ ТЕКСТА АНГЛОЯЗЫЧНОГО КОРОТКОГО РАССКАЗА

М. С. БОГОМОЛЬСКАЯ, И. И. ЗУБОВА

The article deals with the problem of English short story text generation by a computer. The formal model is created with this aim. It includes the following parts: 1) the computer knowledge base with the main and additional keywords of the investigated texts, their logic-semantic and semantic-syntactical formulae and 2) the algorithm. Lexical units necessary for the text creation procedure are presented in the form of the tagged semantic classes

Ключевые слова: база знаний, модель, опорное слово, рассказ, структура, текст

С целью более глубокого изучения речемыслительной деятельности человека, а также законов текстообразования процесс порождения текста можно определенным образом формализовать. Моделирование процесса порождения текста компьютером осуществляется на основе двух видов технологий – шаблонных и лингвистически мотивированных. Шаблонные технологии позволяют порождать тексты, базируясь на готовых шаблонах, созданных человеком. Лингвистически мотивированные системы позволяют порождать тексты с относительно свободным содержанием. Представленная в данной научной работе формальная модель системы автоматического порождения текста англоязычного короткого рассказа является разновидностью последнего вида систем.

В результате автоматического анализа статического содержания текстов пяти коротких рассказов английского писателя А. Мэйли были выделены главные и второстепенные опорные слова каждого текста, задающие его тему. Анализ динамического содержания текстов коротких рассказов позволил описать их логико-семантическую организацию в виде последовательности микротем. Формальное представление семантико-синтаксической организации микротемы осуществлялось с помощью искусственного языка СЕМСИНТ. Для более точного отражения в семантико-синтаксической формуле каждого предложения элементов конкретной ситуации были разработаны довольно объемные семантические подклассы слов знаменательных и служебных частей речи.

Списки главных и второстепенных опорных слов исследованных текстов, их логико-семантические формулы, семантико-синтаксические формулы всех абзацев каждого текста, а также словарь, представляющий собой совокупность семантических подклассов лексических единиц всех текстов составили лингвистическую базу знаний. Она послужила основой для построения формальной модели системы порождения текста англоязычного короткого рассказа. Предложенная модель является

вероятностно-алгоритмической, так как базируется на вероятностном выборе логико-семантической формулы порождаемого текста, семантико-синтаксических формул входящих в него абзацев, а также на определенных лексических, синтаксических и семантических правилах.

Можно предположить, что в порождаемых компьютером текстах в отдельных случаях между некоторыми лексическими единицами будут возникать семантические несоответствия. Это объясняется тем, что вариативность лексических единиц, заданная в перечнях семантических подклассов, в каком-то контексте окажется недопустимой. Поэтому возможность продолжения исследования по данной тематике видится, во-первых, в совершенствовании предложенной формальной модели путем включения в нее результатов анализа большого количества текстов коротких рассказов и построения на их основе универсальной логико-семантической формулы текста данного жанра. Во-вторых, необходимо определить данные о лексической сочетаемости текстовых единиц и включить их в лингвистическую базу знаний, а алгоритмическую модель пополнить точными правилами их выбора в процессе порождения текста. В-третьих, необходимо изучить взаимосвязь сюжета рассказа с последовательностью микротем его абзацев и попытаться ее формализовать. Априори можно предположить, что данная формализация улучшит содержание порождаемого текста.