Т. В. Бусел

ПЕРСПЕКТИВЫ РАЗВИТИЯ НАПРАВЛЕНИЯ АВТОМАТИЧЕСКОЙ ГЕНЕРАЦИИ ТЕКСТОВ НА ОСНОВЕ ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

За последние 5–10 лет в сфере компьютерной лингвистики произошли качественные изменения, предопределившие активное развитие технологии искусственного интеллекта. Искусственный интеллект (ИИ) – это технология не только настоящего, но и будущего, по сути, это наделение компьютеров человеческими способностями, важнейшей из которых является владение языком. Все знания человечества записываются и передаются на естественных языках (ЕЯ), поэтому не удивительно, что способность машин понимать человеческий язык, общаться с людьми, генерировать и переводить тексты находится в фокусе многих научных исследований.

Британский ученый А. Тьюринг был одним из первых, кто всерьез задумался о потенциале ИИ, он верил, что однажды интеллект машин сможет сравняться с человеческим. А. Тьюринг выдвинул идею теста: если в ходе беседы человек не может отличить машину от другого человека, значит, машина достигла уровня человеческого интеллекта. В 2015 г. программа «Eugene Goostman» сумела успешно пройти тест Тьюринга на русском языке.

Технологии обработки ЕЯ, базирующиеся на достижениях ИИ и компьютерной лингвистики, постоянно совершенствуются. Крупными частными компаниями и государственными организациями востребованы системы, моделирующие различные виды речемыслительной деятельности людей, связанные, в первую очередь, с автоматизацией процесса понимания и порождения текстов на ЕЯ. На практике такие системы открывают новые возможности для создания многоязычного информационного контента доступного большому количеству пользователей во всем мире.

Лингвистические ресурсы, необходимые для компьютерной реализации понимания текстов, уже созданы и используются в системах автоматического перевода, индексирования, аннотирования и реферирования. Сложнее обстоит дело с созданием систем, которые позволяют генерировать тексты на ЕЯ. Исследованиями в этой области занимаются ведущие научные центры и университеты мира: Эдинбургский университет в Шотландии, Массачусетский, Стэндфордский и Колумбийский университеты в США, Монреальский университет в Канаде, а также МГУ имени М. В. Ломоносова и Российский НИИ искусственного интеллекта.

Основные трудности, с которыми сталкиваются ученые и разработчики систем, моделирующих интеллектуальную деятельность человека по порождению текстов, как правило, носят лингвистический характер. Их решение связано с формализацией грамматики и лексических описаний, выявлением правил и процедур преобразования семантических структур в естественно-языковую форму, с исследованием лингвистических характеристик текстов и риторических приемов организации их содержания, а также языковых средств выражения связности текста и рядом других довольно сложных задач.

Автоматическая генерация текстов может быть осуществлена в рамках различных подходов (М. В. Болдасов, Р. Дейл, Дж. Лестер, Э. Райтер, Р. Рубинофф, Е. Г. Соколова, Э. Хови), которые основаны на использовании шаблонов (готовых текстовых фрагментов) и лингвистических знаний. Недостаток первого подхода заключается в том, что он не может быть использован для моделирования содержания и структуры порождаемого текста, поскольку предполагает алгоритмическую подстановку готовых текстовых фрагментов (предложений, которые обычно имеют цифровой идентификатор и содержатся в базе данных в виде списка) в заданные позиции в документе без какой-либо их дополнительной обработки.

Второй подход основан на использовании лингвистических правил, которые применяются для эксплицитного описания знаний о структуре и содержании генерируемого текста, а также знаний, которые позволяют выразить это содержание языковыми средствами. Архитектура системы автоматической генерации текстов, основанная на данном подходе, как правило, состоит из нескольких модулей, обеспечивающих координацию всех уровней лингвистического описания языка. Модуль планирования содержания текста включает два подмодуля, работающих в тесном взаимодействии. Первый подмодуль определяет, какая информация будет участвовать в генерируемом тексте, а второй – порядок следования информации в документе. Результатом работы данного модуля является логико-семантическая модель текста.

В модуле языкового оформления текста происходит выбор лексических и грамматических средств, представленных в базе знаний системы, которые необходимо использовать, чтобы выразить содержание, сформированное в первом модуле. На данном этапе генерации осуществляется выбор лексических единиц и их грамматических форм, выполняется морфологическое согласование между членами грамматических групп, а также проверка соответствия выбранных слов структурно-семантическим и лексико-семантическим правилам сочетаемости. Таким образом, данный модуль переводит логико-семантическое представление, построенное предыдущим модулем, в текст на ЕЯ.

Наиболее эффективным в настоящее время является подход к решению задачи автоматической генерации текста, который базируется на нейросетях и предполагает использование алгоритмов машинного обучения. Нейросети, которые обучаются на огромных объемах данных, называются языковыми моделями. Большие языковые модели – одно из наиболее востребованных и интересных направлений развития ИИ. Принцип работы таких моделей основан на определении вероятностного сочетания слов и их значений в заданном контексте с использованием определенных алгоритмов.

Ведущие позиции среди современных языковых моделей занимает многоязычная языковая модель BLOOM (BigScience Large Open-science Open-access Multilingual Language Model), созданная в 2022 г. специалистами из Французского национального центра научных исследований (French National Center for Scientific Research). Данная модель обладает более чем 176 мил-

лиардами параметров, которые определяют, как ИИ выполняет вычислительную задачу. Чем больше таких параметров включает в себя языковая модель, тем более сложные задачи ей под силу.

Модель BLOOM обучали на огромном количестве языков, поэтому она может отвечать на вопросы, вести диалог, а также обрабатывать, генерировать и перефразировать тексты на 46 естественных языках, включая французский, испанский, русский, арабский, английский, вьетнамский, китайский, индонезийский, каталанский, хинди, бенгали и т.д. Среди основных проблем в работе с языковой моделью эксперты выделяют сложность обновления базы фактов, на которых обучена модель, и отсутствие ссылок на источники.

Модель BLOOM создана по принципам «открытой науки». Ее можно использовать в образовательных и научных целях, создавать уникальный качественный контент за считанные минуты: статьи, рефераты, сочинения, научные работы и т.д. Следует отметить, что возможности ИИ вызывают справедливые опасения у научного сообщества, поскольку доказать плагиат в таком случае – довольно сложная задача. Как отмечают эксперты, существуют несколько потенциальных способов использования BLOOM в образовании:

1. В качестве инструмента для изучения языка. Генератор текстов может быть использован для того, чтобы помочь студентам практиковать свои языковые навыки, создавая тексты на определенном языке или в определенном стиле.

2. Как способ изучения искусственного интеллекта. Студенты могут использовать генератор текстов как способ узнать о работе нейронных сетей и о том, как они могут эффективно применяться для автоматической обработки ЕЯ (Natural Language Processing).

Разработчики языковой модели BLOOM отмечают, что внедрение ИИ значительно расширило область применения технологии автоматической генерации текста и синтеза речи. В недалеком будущем синтез станет применяться для автоматического создания новостных сюжетов, озвучивания фильмов, игр и интерактивных образовательных курсов. Каждый сможет создать цифровую копию своего голоса и свободно общаться на различных языках. Таким образом, языковая модель BLOOM может помочь пользователям в решении различных лингвистических задач, а также сделать процесс обучения интереснее и эффективнее.