**Т. В. Бусел**

# СОВРЕМЕННЫЕ ИНСТРУМЕНТЫ ЛИНГВИСТА И ПЕРЕВОДЧИКА: БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ НА ОСНОВЕ ТЕХНОЛОГИИ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Искусственный интеллект (ИИ) – это область науки, которая стремительно развивается и постоянно ставит новые задачи и вызовы перед человечеством. Британский ученый А. Тьюринг был одним из первых, кто задумался о возможности создания ИИ и разработал теоретическую и философскую концепцию этой идеи. Ученые полагают, что ИИ обладает огромным потенциалом для достижения благих целей: от открытия новых областей научных исследований до улучшения качества образования. Если будут реализованы все потенциальные возможности, открывающиеся благодаря использованию инструментов ИИ в сфере науки, образования и других областей, представляющих общественный интерес, это может сыграть важную роль в развитии общества, а также изменить его к лучшему.

ИИ – это технология не только настоящего, но и будущего, по сути, это наделение компьютеров человеческими способностями, важнейшей из которых является владение языком. Язык – это память человечества, хранящая его знания, культуру и историю. Поэтому не удивительно, что построение ИИ моделей, описывающих функционирование человеческого языка и его связи с мышлением, а также позволяющих создавать универсальные интеллектуальные системы, обладающие знаниями и способные принимать решения, находится в фокусе современных научных исследований.

ИИ модели бывают разных видов, каждый из которых обладает уникальной архитектурой и подходом к пониманию и генерации естественного языка. Эти модели значительно эволюционировали со временем, пройдя путь от традиционных статистических методов к более сложным архитектурам на основе нейронных сетей, разработкой которых занимаются ведущие научные центры и университеты мира: Пекинская академия искусственного интеллекта, Кембриджский и Оксфордский университеты, на базе которых был создан Центр по изучению искусственного интеллекта и будущего человечества, Институт перспективных исследований проблем искусственного интеллекта и интеллектуальных систем МГУ имени М. В. Ломоносова, Стэндфордский университет и Объединенный институт проблем информатики НАН Беларуси.

Первые модели ИИ строились под конкретные языки, однако в настоящее время одним из основных требований к подобному программному обеспечению является мультиязычность. По данным, опубликованным в научном журнале New Scientist, большая языковая модель, получившая название No Language Left Behind (NLLB), может осуществлять перевод с 204 языков, включая редкие языки, такие как ачехский, фриульский, урду, а также языки коренных народов Африки и Австралии. Несмотря на свое

название, модель NLLB охватывает лишь незначительную часть из почти 7000 языков, существующих во всем мире. Модель NLLB по уровню качества перевода в среднем на 44 % превосходит ранее предлагаемые исследовательские системы на основе машинного обучения при использовании метрик BLEU, сравнивающих машинный перевод с эталонным человеческим переводом.

Развитие технологий ИИ способствует появлению многоязычных моделей, которые позволяют распознавать и синхронно переводить человеческую речь, а также выполнять аудиовизуальное дублирование и языковую локализацию. На основе NLLB была создана языковая модель SeamlessM4T, которая способна клонировать оригинальный голос с сохранением стиля, акцента и интонаций на 100 языков. Модель также позволяет выполнять синхронный перевод и поддерживает редкие языки, в том числе и наречие хоккиен (тайваньский язык), не имеющее собственной письменности. По данным, опубликованным в научном журнале Nature, благодаря объединению нескольких задач перевода в единую многогранную модель технология оптимизирует процесс перевода и значительно повышает эффективность.

Одним из ключевых факторов, который делает большие языковые модели уникальными, оказывается их способность обучаться. Модели, такие как NLLB и SeamlessM4T, как правило, обучаются на огромных объемах разнообразных текстовых данных, таких как корпусы текстов, книги, статьи, веб-страницы и т. д. В процессе обучения модель “учится понимать” языковые закономерности, структуру предложений и контекст. Обучение больших языковых моделей – это сложный и ресурсоемкий процесс, требующий комбинации передовых алгоритмов машинного обучения и больших объемов данных.

Большие языковые модели, как правило, использует трансферное обучение, это означает, что они могут применять полученные знания из одной задачи, чтобы обучиться выполнению другой задачи. Например, модель NLLB, обученная переводить с английского на испанский язык, может использовать полученные знания для перевода с английского на немецкий язык.

Обобщая исследования и разработки в сфере ИИ, можно выделить следующие наиболее перспективные задачи, которые могут быть решены с помощью больших языковых моделей:

- обработка естественного языка, поиск и извлечение информации из текстов;
- распознавание и синтез речи;
- автоматический перевод;
- анализ тональности текстов;
- исследование и сохранение языков, включая редкие и исчезающие языки (большие языковые модели способны генерировать тексты и переводы на различных языках, что помогает документировать и изучать культурное наследие);

• создание различных приложений, связанных с языком, таких как диалоговые системы, виртуальные помощники, системы машинного перевода и т. д.

В сфере образования большие языковые модели и специальные программы, созданные на их основе, такие как, например, Duolingo Max на базе GPT-4 становятся незаменимыми помощниками, как для преподавателей, так и для студентов. Они способны предлагать индивидуальные учебные программы и онлайн курсы, накапливать актуальную информацию и предоставлять доступ к обширным знаниям, тем самым повышая эффективность процесса обучения.

Большие языковые модели являются фундаментом новой цифровой реальности, в которой языки и технологии идут рука об руку. Большинство систем ИИ разработано для языков с высоким уровнем ресурсов, таких как английский, испанский и китайский, что создает серьезный языковой разрыв и лишает многих, в том числе белорусов, доступа к передовым технологиям на их родном языке. В этом контексте актуальной задачей в Республике Беларусь становится создание и развитие национальной языковой модели, способной поддерживать и продвигать белорусский язык в цифровом пространстве.

Создание ИИ технологий и систем является одним из приоритетных направлений развития научной, научно-технической и инновационной деятельности в Республике Беларусь. В связи с этим, важным, с практической точки зрения, является проведение фундаментальных и прикладных научных исследований, направленных на создание национальной языковой модели, которая сможет эффективно обрабатывать белорусский язык, учитывая культурные, лингвистические и деловые особенности страны, и предоставлять технологическую независимость в этой области.

Создание национальной большой языковой модели открывает новые возможности:

• сохранение и распространение белорусского языка, его интеграция в современные цифровые технологии для развития общества, экономики и науки в Республике Беларусь;

• создание национальной базы данных, которая станет основой для дальнейших исследований и разработок в сфере ИИ;

• развитие образования – разработка обучающих систем и платформ онлайн-обучения с использованием ИИ, которые позволят создавать персонализированные учебные программы для белорусских и иностранных студентов, учитывающие их культурные особенности и предпочтения.

ИИ помогает ученым, лингвистам и переводчикам создавать новые горизонты для исследований и инноваций. Поддерживая развитие ИИ, создавая национальную языковую модель, общество получает уникальную возможность сохранить свое языковое и культурное наследие и ускорить технологическое развитие.