Борис Орехов: «Теперь мы пишем не как хотим, а как это подсказывают наши компьютеры»

Уфимец Борис Орехов, кандидат филологических наук, ныне доцент московской Высшей школы экономики, рассказал RBtoday, как компьютерная лингвистика может приносить деньги и создавать полезные коммерческие продукты, почему башкирский язык распространен в Интернете и что мы теряем, когда язык умирает.

«КОМПЬЮТЕР ВЕДЬ ДОВОЛЬНО ГЛУПАЯ ШТУКА»

- В компьютерной лингвистике – одной из сфер ваших профессиональных интересов – программирование сочетается с филологической наукой. Это равноправные половины или что-то доминирует?

- Компьютерная лингвистика занимается созданием технологий, которые позволяют обработать большое количество текстов, извлечь из них информацию или перестроить, сделать сложный текст простым. Компьютерный лингвист не пишет код и не исследует язык. Он пытается понять, как сложные языковые схемы можно сделать проще, понятней для компьютера. Лингвист привык работать с языковым материалом, поэтому ему легче улавливать тенденции и формализовать их, дело программиста – написать код, который эти схемы встраивал бы в компьютерный продукт. Еще важно провести границы между наукой и инженерией. Наука думает об абстрактном и разрабатывает модели мироустройства, а инженерия пытается эти закономерности применить на практике. Так, гениальные физики изучали устройство атома, а инженеры старались из этого знания сделать атомную электростанцию, которая производит электричество для конечного пользователя. Конечно, существует количественная лингвистика, цель которой не коммерческий продукт, а понимание как устроен мир, то есть такой андронный коллайдер. Но лингвистика компьютерная – больше инженерная область. Главное в ней – формализация, представление о том, как все многообразие слов преобразовать в простую и понятную компьютеру форму. Компьютер ведь довольно глупая штука. Мы сейчас его идеализируем и пытаемся поставить на пьедестал, но на самом деле он мало что может.

- Эта область исследований появилась с первыми компьютерами. Как она изменилась за полвека?

- Одно из главных условий компьютерной лингвистики – много текстов в электронном виде. В середине века компьютеры занимали огромные помещения, были ламповыми, а жесткие диски на несколько сотен мегабайт стоили запредельные суммы, которые были не под силу индивидуальным пользователям. Сейчас хранение больших текстов в «цифре» не проблема – на телефоне, на флэшке можно легко уместить собрание сочинений. Поэтому хоть и сами идеи были заложены достаточно давно, их начали всерьез применять только в последние 15-20 лет. Эти достижения уже вошли в нашу жизнь, мы незаметно для себя ими пользуемся, когда, например, вводим слова в мессенджере, пользуемся поисковиками или переводчиками в Интернете.

- Результат работы компьютерных лингвистов – коммерческий продукт. Кто ставит задачи перед разработчиками?

- В мире появился огромный поток текстовой информации и прочитать его человек не способен, даже если нанять тысячу работников, что уже дорого и нерентабельно. Проще заставить компьютер сделать это, чтобы он извлек полезную информацию, скажем, об упоминании определенных людей или брендов. Обычно такие задачи ставят менеджеры крупных компаний. Например, телекоммуникационная компания заинтересована в своем имидже среди клиентов. Она заказывает компьютерно-лингвистической конторе мониторинг соцсетей, где каждую секунду появляется по несколько сообщений. Компьютер выискивает необходимые упоминания и пытается понять, хорошо здесь сказано о бренде или плохо, а если плохо, то почему. Это способ улучшения имиджа, в конечном счете – роста продаж. Любой текст устроен довольно сложно: необязательно все выражено теми словами, которыми он написан. Например, мы ищем что-то про королей, а в выдаче видим тексты с упоминанием слова «монарх». То есть компьютер уже умеет различать слова, похожие по смыслу, а не по внешнему виду, и группировать такие тексты тематически. Еще один пример связан с коррекцией орфографии – к слову о Тотальном диктанте, который пройдет по всей России. Появился интересный тренд: чем дальше, тем грамотнее люди пишут в соцсетях. Есть социологи, которые пытаются связать это с эффектом Флинна, согласно которому у каждого нового поколения балл по тесту на IQ выше, чем у предыдущего поколения. На мой взгляд, все объясняется проще – теперь мы пишем не как хотим, а как это подсказывают наши компьютеры. В телефоны, в соцсети, в браузеры встраивается проверка орфографии и автозамена на более правильное написание. То есть это не то, как человек на самом деле умеет писать, а то, как его исправляет компьютер. Это тоже определенные алгоритмы, довольно понятные. И куда их встраивать – зависит от конкретной прикладной области.

- В конечном счете, эти разработки помогают созданию искусственного интеллекта?

- Искусственный интеллект – все-таки отдельная область со своими задачами, хотя тоже глубоко инженерная. Я бы не стал воспринимать тех, кто занимается развитием искусственного интеллекта, людьми в белых халатах, засевших в лабораториях в попытке изобрести нечто, что начнет понимать искусство, юмор, глубину культуры. Это вполне практическая задача, в соответствии с которой мы заставляем компьютер решать человеческие задачи интеллектуального характера. Например, распознавание лиц на фото – чаще всего современные задачи ИИ связаны больше с обработкой изображений и видео, чем с текстами.

- Компьютер пока не научился считывать иронию и учитывать настроение текста. Машина когда-нибудь сможет понять человека?

- Настроение, точнее анализ тональности, как раз прочитывается компьютером – с меньшей погрешностью, чем при распознавании иронии. Настроение – важная

штука. Когда пишешь пост про своего оператора мобильной связи, компьютер должен понять, с каким настроением он написан. В этой области все неплохо – эти задачи стали решать нейросети, которые выдают существенно лучший результат, чем у нас был лет 5-10 назад. Понятно, что нейронные сети не способны решить все стопроцентно, но, кажется, в анализе тональности прогресс еще будет. С иронией все намного сложнее. Это действительно что-то человеческое – даже слишком человеческое, как сказал бы Ницше. Здесь прогресс не такой большой, и специалисты делятся на оптимистов и пессимистов. Пессимисты говорят, что мы упремся в потолок и дальше него не пойдем, оптимисты – что при нынешней скорости развития технологий нас ждет глобальный прорыв. Удастся ли решить эту проблему? Тут нужно обратиться к визионерам, которые чувствуют будущее и дальнейшее развитие рынка – вроде Илона Маска.

- Чем вам интересна эта область исследований? Что вы хотите с ее помощью открыть, разработать?

- Я не типичный компьютерный лингвист. Они чаще всего берут технологию, которая работает на 98%, и пытаются ее улучшить, чтобы она работала на 98.5%. Я беру существующие, хорошо работающие вещи и применяю их там, где они не применялись до сих пор. В частности, один из моих интересов – создавать работающие компьютерные инструменты для языков, не очень интересных крупному бизнесу, который готов вкладываться в продукты для английского, французского, испанского, иногда для русского. Поскольку нет рынка, с которого можно было бы получить деньги для интеграции удмуртского, бурятского, башкирского языка, это становится делом университетских ученых. Ну и, конечно, меня интересуют цифровые гуманитарные науки, то есть как можно компьютерные инструменты применить к литературе. Получаются довольно интересные результаты.

- Эти результаты интересны широкой аудитории или это больше чтение для специалистов?

- Трудно сказать. Насколько я вижу, сейчас интерес к науке вообще существует и он гораздо выше, чем 10 лет назад. Количество научно-популярных ресурсов выросло – сходу могу назвать дюжину таких. Поскольку любую науку можно изложить популярно, то запрос на подобные исследования в разных областях довольно высок. Надеюсь, не только в пределах Садового кольца. Речь о том, что человеку всегда – все 40 000 лет существования в виде homo sapiens – было интересно, как устроен мир вокруг него. Язык и искусство тоже как-то сделаны, и до сих пор не очень понятно как. Несмотря на то, что лингвистика существует несколько тысячелетий, а сейчас у нас компьютеры под рукой, просто описать систему языка по-прежнему нельзя.

«МЫ ХОРОНИМ УСИЛИЯ ЛУЧШИХ ПРЕДСТАВИТЕЛЕЙ ЧЕЛОВЕЧЕСТВА»

- Новых языков почти нет, а те, что есть, исчезают. Почему это происходит?

- Новые языки еще иногда образуются, хотя это происходит не так, как было, скажем, с русским. Например, коренной народ, живущий на острове, в силу обстоятельств коммуницирует с колонизаторами. В процессе взаимодействия они вырабатывают так называемый «пиджин» – язык, который является особым соединением двух исходных. Кроме того, те или иные диалекты могут приобрести статус полноправного языка. То есть потенция возникновения новых языков все-таки есть. Умирание языков – гораздо более сильная тенденция и зависит она от того, считает ли носитель языка его престижным. Дело, конечно, не в материальной выгоде, а в символическом капитале. Если в коллективе теряется это понимание, считается, что другой язык выгоднее и удобнее, он социально востребован, тогда родной язык уходит. Крупные языки дают довольно много общественных, финансовых и социальных преференций. У носителя появляются дивиденды благодаря тому, что он говорит на русском, а не на языке маленькой деревни в Забайкалье. Почему, собственно, родители озабочены тем, чтобы их чадо обязательно учило иностранный? Речь как раз о тех самых дивидендах, благодаря которым ребенок потом не потеряется. Трудно себе представить родителей, живущих в Псковской области, которые говорят сыну, чтобы тот обязательно выучил калмыцкий, потому что без него не прожить. В итоге все приводит к сокращению числа носителей и к общей печальной тенденции.

- Что мы теряем в случае смерти языка?

- Прежде всего, информацию – самое ценное, что есть в современном обществе, и эта информация очень разноплановая. Язык составляет рамку, в которой может быть выражена человеческая мысль. Если мы попытаемся ее выразить на другом языке, то, как свидетельствуют исследования гуманитариев, это будет уже другая мысль. Другими словами, полный и адекватный перевод с одного языка на другой невозможен. Возьмем древнегреческую философию: насколько хорошо мы понимаем того же Платона? Оказывается, он специально подбирает такие слова, которые помогают ему выразить мысль, и они не равнозначны русскоязычному переводу. Это касается не только слов, но и целых конструкций, которые сейчас воспринимаются иначе, чем в древние времена. Например, в «Слове о полку Игореве» есть выражение «растекаться мыслью по древу». Сейчас это значит много болтать, не говорить ничего содержательного. Но ведь в «Слове» это означает другое. «Растекаться» – глагол движения, который не был связан напрямую с движением жидкости, а означал передвижение любого объекта. У Пушкина в «Анчаре»: «…и тот послушно в путь потек». Пушкин не имел в виду, что персонаж растворился в какой-то жидкости и улетучился, он просто начал двигаться. При исчезновении языка мы теряем представление, что на самом деле было сказано. Человеческая мысль, которая формировалась тысячелетиями и которая составляет основу нашей культуры, оказывается искажена. Древнегреческий язык исчез, его знают по каким-то отрывкам, в итоге Платона мы понимаем не совсем правильно.

- Если оказывается искажена философия, что говорить о поэзии.

- В стихе непросто выразить поэтическую идею, авторский замысел, для этого нужно вложить очень много усилий. Известная история про «Медного всадника»: Пушкин, чтобы найти нужный эпитет, постоянно вычеркивает слова и дописывает новые – и так делает до 12 раз. То есть он чувствует, что конкретное слово – не то, что нужно. И если мы теряем язык, то теряем все эти усилия, культурные и интеллектуальные, которые были вложены человеком в попытке выразить идею. Текст зачастую выстрадан с точки зрения стилистики и содержания, и получается, что мы хороним усилия лучших представителей человечества, которые на этом языке говорили. Еще одна проблема лежит в профессиональной плоскости. В языке хранится много исторической информации. Мы можем сопоставлять языки между собой и лучше понимать, как они выглядели тысячелетия назад. Если они умирают, то у нас теряется возможность такой реконструкции. Не говоря уже о том, что разные языки по-разному выражают разные идеи – пространственной ориентации, цвета и т.д. При этом язык на данном этапе все еще нельзя описать полностью, мы не можем зафиксировать его так, чтобы узнать про него все и не потерять информацию о том, как он устроен, после его исчезновения.

- Как предотвратить умирание языков?

- Видимо, повышать их престиж. Например, административными способами, которые есть у государства. Преподавание языка в школах и вузах – это, безусловно, знак престижности, востребованности. Также на повышение статуса работает то, что делает язык более современным для его носителей, то есть формирует более современный имидж – на нем появляются компьютерные игры, он является интерфейсным языком соцсетей, сайтов. В частности, некоторые крупные российские интернет-компании вводят в свои интерфейсы национальные языки. До малых языков у них дело не доходит, но все же. Думаю, такие проекты не приносят им прибыли – они имиджевые, но очень хорошо, что они существуют.

- Сейчас среди малых языков в России лидеры по распространенности – татарский, башкирский, якутский и удмуртский...

- Еще чеченский. Но стоит оговориться: есть две разные сферы распространения – офлайн и онлайн. Если мы говорим про живых людей, которые ходят по улицам, то да, чеченский – один из лидеров. Если мы говорим про тексты, которые пишутся в Интернете, то у чеченского довольно низкий старт в Сети в силу социоисторических причин, которые мы наблюдали в 90-х и начале нулевых.

- Почему именно эти языки получили такое распространение в Интернете?

- Как правило, за любой такой сетевой активностью стоит активное интернет-сообщество. Говорят же, что революцию нельзя провести, если нет сплоченной группы активистов – просто даже при поддержке населения не найдется тех, кто будет осуществлять захват власти. Примерно так же и в Сети. При этом витальность языка в Интернете мало зависит от того, сколько человек говорит на этом языке в офлайне. Важнее то, сложилось ли вокруг него активное сетевое сообщество, действующее на престижных интернет-площадках.

- В каком состоянии находится русский язык? Он продолжает развиваться?

- Есть разные показатели. Много ли на нем говорит людей, становится ли их больше? Нет, количество говорящих на русском языке сокращается, он становится менее востребованным. Очевидно, в силу утраты геополитического влияния страны, для которой русский язык был главным – потеряна Прибалтика, все меньше людей говорит на русском в Средней Азии. Что касается свойств, которые находятся внутри языковых процессов, то русский чувствует себя ничуть не хуже, чем английский. Он по-прежнему умеет образовывать новые интересные слова и новые значения. Например, прекрасные, хотя и уже устаревающие слова «распил» и «откат». Они были в русском языке столетия назад: распил – специальный термин, связанный с лесозаготовкой, откат – смещение артиллерийского орудия после выстрела, если оно установлено на колесах. В какой-то момент русский язык решил, что эти слова не очень широко распространенные и их можно использовать для новых значений и новых реалий. Что и произошло примерно в начале нулевых. Они стали употребляться в новом смысле и это, безусловно, один из маркеров того, что язык развивается и в нем происходят живые процессы.

Андрей КОРОЛЁВ