статья // №24 (152), 24 июня 2010

Машина в поисках смысла

Ольга Андреева, Григорий Тарасевич

Полвека назад кибернетики были уверены, что машины скоро научатся переводить Пушкина и Шекспира. Однако скоро не получилось. Машинные переводчики уже способны на многое. Но они до сих пор не умеют главного — понимать смысл того, что они переводят. Решая эту проблему, наука наталкивается на вечные вопросы: что такое язык, что такое жизнь и как мы ухитряемся понимать друг друга

Большая аудитория РГГУ. На кафедре знаменитый лингвист Игорь Мельчук. Бывший советский ученый, а ныне канадский профессор специально приехал, чтобы прочитать лекцию с лаконичным названием «Смысл». За полтора часа огромная доска покрывается кружочками, стрелочками и прочими значками.

В конце лекции из зала раздается:

— Простите, так что же такое смысл?

Мельчук пожимает плечами и указывает на исписанную вдоль и поперек доску:

— Как? Разве что-то не ясно? Вот это все и есть смысл!..

Все что нужно — это взломать код

Эволюция машинного перевода — история того, как идея, казавшаяся поначалу очень простой, в процессе исполнения превращается в неподъемную задачу. Что такое машинный переводчик? Да просто черный ящик, внутрь которого поступает русская

фраза, а выходит английская — того же содержания. Если вы сами знаете иностранный язык, роль черного ящика выполняет ваша голова.

«У меня перед глазами текст, написанный по-русски, но я собираюсь сделать вид, что на самом деле он написан по-английски и закодирован при помощи довольно странных знаков. Все, что мне нужно, — это взломать код, чтобы извлечь информацию», — писал в конце сороковых годов Уоррен Уивер, директор отделения естественных наук Рокфеллеровского фонда. С этой нехитрой мысли все и началось. Перевод казался прелестной игрушкой, на которой можно было продемонстрировать мощь электронных технологий.

Оная демонстрация состоялась 7 января 1954 года и вошла в историю под названием Джорджтаунского эксперимента. Специалисты из одноименного университета совместно с компанией IBM впервые в мире автоматически перевели 49 предложений с русского языка на английский. Научная общественность была потрясена. США, а за ними и многие другие кинулись финансировать соответствующие проекты. Но…

— Знаете, что такое Джорджтаунский эксперимент? — спрашивает один из создателей современной системы машинного перевода «Кросслейтор» Эдуард Клышинский. — Чистое жульничество. Представьте себе словарь из 250 слов и аналитический аппарат из 6 правил. Простейшие фразы, соответствующим образом составленные, вы переведете как по маслу. Но возьмите словарь в 1000 слов. Это будет сложнее не в 4 раза. По мере увеличения словаря сложности будут нарастать экспоненциально. За каким-то порогом вы вообще не получите результата. Текста просто не будет.

Машина в Джорджтауне понятия не имела, где в предложении подлежащее, а где сказуемое. Напрочь не различала ни одной формы слова и ни одной части речи. Фразы составлялись так, чтобы достаточно было простых соответствий: «мама» — «mother». Собственно, в этом и состояло невинное кибернетическое жульничество.

Проклятый Джон и его игрушки

Спустя всего шесть лет после Джорджтаунского эксперимента машинный перевод был торжественно похоронен. Убила его простенькая фраза: «John was looking for his toy box. Finally he found it. The box was in the pen». Ее правильный русский перевод звучит так: «Джон искал свою коробку с игрушками. Наконец он ее нашел. Коробка была в манеже».

Автор фразы, американский философ Иегошуа Бар-Хиллел, заявил, что для слова «pen» («ручка», но оно же и «детский манеж») ни один электронный переводчик никогда не сможет подобрать точный аналог на другом языке. Выбор между «ручкой» и «манежем» можно сделать только имея определенную картину мира, которой у машины нет. По мнению Бар-Хиллела, этот факт закрывал тему электронного перевода навечно. Кстати, до сих пор ни один онлайн-переводчик перевести эту фразу не в состоянии. Мы проверяли.

Окончательно добила первые переводчики так называемая Черная книга машинного перевода — опубликованный в 1966 году доклад Комитета по прикладной лингвистике Национальной академии наук США. Группа экспертов констатировала принципиальную невозможность машинного перевода и советовала работы в этом направлении прикрыть. Что и было сделано.

Поймать муху на Луне

Черное десятилетие машинного перевода на Западе длилось с середины 60−х по середину 70−х. Россию от той же участи спас железный занавес. Более того, у нас для структурной лингвистики наступил «серебряный век». Разрабатывались собственные системы машинного перевода, открывались специализированные кафедры и лаборатории. Недаром один из персонажей «Попытки к бегству» братьев Стругацких напевал:

Воет ветер дальних странствий,
Раздается жуткий свист —
Это вышел в Подпространство
Структуральнейший лингвист.

Представитель языкознания встал вровень с космофизиками, олицетворяя собой технологическую утопию. В том же романе эта утопия показана в действии: с помощью «мнемокристалла» можно было запросто понимать даже язык инопланетянина.

Впрочем, действительность быстро возвращала на землю. Классическая лингвистика на тот момент почти ничего не могла предложить кибернетикам, кроме общих принципов. Никому и в голову не приходило составлять, допустим, точные перечни всех синтаксических конструкций, возможных на том или ином языке. А ведь тупой машине нужно было именно это. Кстати, тогда ЭВМ занимала несколько этажей, работала на перфокартах, а в очередь на час работы записывались примерно за месяц.

— В свое время у нас был такой тост: за мечту, которая никогда не сбывается, — говорит Эдуард Клышинский. — Понимаете, сделать переводчик — примерно то же самое, что поймать муху на той стороне Луны. Для этого надо туда прилететь, создать условия, чтобы муха могла там жить, потом поймать и привезти обратно.

По счастью, наука редко отказывается от Мечты. Так что мух на той стороне Луны ученые еще половят.

— Хотя бы для того, чтобы на этой Луне порыбачить, — добавляет Клышинский.

Статистика вместо понимания

— Я представляю компанию «Яндекс», — скромно говорит юноша в скромной майке. — Я узнал, что еду на конференцию «Диалог», уже после того, как вышел из дома, поэтому прошу простить меня за внешний вид. Давайте переведу, что у меня написано на майке: «“Да брось ты свой компьютер, пойдем погуляем”, — говорит “Гугл”».

Молодой человек поворачивается спиной, и аудитория, состоящая из лингвистов и математиков, читает: «Fuck GOOGLe». Как много, однако, теряется при переводе!

— Я хочу сказать, что все решения исходят из конкретной задачи… — продолжает юноша.

На ежегодном «Диалоге», конференции по проблемам компьютерной лингвистики, молодой человек из «Яндекса» представляет коммерческие структуры. Его маечка на фоне клетчатых рубашек научных сотрудников напоминает о том, что за все надо платить.

Именно конкретные задачи толкали машинный перевод вперед, несмотря на все концептуальные преграды. Американским инженерам нужно было переводить тонны советской технической документации — лингвисты получали финансирование. В начале

90−х малограмотные российские бизнесмены хотели вести дела с иностранцами — дискеты с системой «Промт» раскупались по цене «жигулей». Миллионы пользователей интернета не владеют английским — онлайн-переводчики могут стать выгодной опцией.

Зовущий гулять Google был одним из первых, кто соединил поисковик с переводчиком. Несколько нажатий мышью — и кореец может читать французский сайт, немец — американский, араб — русский и так далее.

Качество средненькое, но суть уловить можно. Беда в том, что за этим переводчиком слишком мало науки. Он относится к новому классу — статистический перевод. Принцип прост: зачем переводить заново то, что уже когда-то было переведено?

— Есть хорошо развитые языки — скажем, английский и русский, — для которых существует огромное количество параллельных переводов — романов, технической документации и прочего. Дальше чисто математическими методами система находит в этом море текстов тот, который статистически ближе переводимому фрагменту, — объясняет лингвист Леонид Иомдин.

Допустим, у вас в базе данных есть «Война и мир», инструкция по использованию стиральной машины и их переводы на английский язык. Нужно разобраться с фразой: «После минутного молчания она начала снимать свою шубу из искусственного меха». Перевод первой части фразы можно найти у Толстого, второй — в инструкции. Если что-то не так, разработчики или даже сами пользователи могут предложить лучший вариант перевода. Поэтому кажется, что система с каждым днем становится все более умной.

— За счет того, что статистические системы выдают вполне приемлемое качество, появляется иллюзия, что проблема вот-вот будет решена, — печально говорит Клышинский. — Но статистика — это не перевод вообще.

Например, Google уверенно переводил название «ул. Владимирская» как «sent (святой) NASDAQ». Почему? Ответ как в старом мультике — так посчитали. Чистая статистика и никакой попытки понять смысл.

От текста к смыслу и обратно

Главную проблему автоматического перевода можно передать одним коротким словом «смысл». Надо научить машину понимать вводимую в нее информацию. Тогда Джордж будет находить свои игрушки в манеже, а президент Bush не окажется кустарником.

— Мы должны начать с того, что такое язык, — говорит академик Юрий Апресян, лингвист с мировым именем, уже полвека занимающийся семантической, то есть смысловой, природой слова. — Все попытки рассматривать язык как код провалились. Но если язык не код, тогда что? Мы имеем в голове некую мысль, находим для нее адекватное языковое выражение, а тот, кто нас слушает, совершает обратную операцию — обращает языковую форму в смысл. Так язык выступает в качестве посредника во взаимном понимании. Но я не занимаюсь электронным переводом — я пытаюсь построить универсальную модель языка.

Модель языка должна работать по принципу: на входе — смысл, а на выходе — текст. Или наоборот.

— Сделать это не так легко, — поясняет коллега Апресяна лингвист Леонид Иомдин. — Прежде всего потому, что текст можно увидеть, услышать, прочесть, а смысл ненаблюдаем: он в голове, и в общем-то про него ничего неизвестно.

Условно говоря, между текстами на английском и русском должно стоять нечто промежуточное — так сказать, язык без языка. Этот посредник получил название семантического представления, или метаязыка. В нем только чистый смысл.

Состоять метаязык должен уже не из слов, но из семантических первоэлементов, неделимых единиц смысла. Юрий Апресян был одним из тех, кто эти элементы впервые описал и дал им название — семантические кварки.

— Это такие элементы, для которых нет соответствий в словах языка, — объясняет Иомдин. — Ну, например, возьмем фразу: «Я стою перед шкафом». Ее смысл зависит от ориентации двух объектов относительно друг друга. Вот эта «лицевость», или «фронтальность», — это и будет семантический кварк. Представить это словом невозможно. А кварком — вполне.

Любимое занятие математических лингвистов — формализовывать все неформализуемое. Отношения реального мира можно загнать в схему, где есть агент (тот, кто делает), причина (почему делает), время (когда делает) и так далее. Даже если перед нами инопланетянин, смысл его стояния перед инопланетным шкафом будет выражаться все той же универсальной «лицевостью».

Второе с половиной поколение

— Вот мы все анализируем, анализируем, бесконечно анализируем, и это все еще Shallow!

Все тот же «Диалог». Интеллигентная структурная лингвистка средних лет в отчаянии заламывает руки. Shallow — это поверхностный уровень синтаксического анализа текста в процессе машинного перевода. За ним должен последовать Deep, то есть глубинный уровень, выводящий на понимание смысла. Должен, но пока не следует…

— Нет, подождите! — В процесс вторгается не менее интеллигентный информатик. — Вот у нас прошла морфология…

Информатик делает изящный шаг вперед:

— Вот пошел поверхностный синтаксический уровень, еще один шаг… — Но лингвистка не выдерживает:

— Вы так от нас уйдете! Когда же начнется Deep?..

Увы. Реально работающего, всеобъемлющего семантического представления до сих пор нет. Существуют только уровни анализа, которые к нему приближают. Первое поколение переводчиков — это перевод на уровне морфологических структур. Второе поколение — это синтаксические структуры. Третье поколение переводчика по идее должно считывать чистый смысл текста, что сделает возможным перевод с любого языка на любой. Но это — мечта. Сейчас Апресян с коллегами разрабатывает систему автоматического перевода «Этап-3», которую условно называют «системой второго с половиной поколения».

— В нашем понимании текста, может быть, мы проникли чуть глубже, чем другие переводчики такого же типа. Но добраться до чистого смысла пока не получается. Так до сих пор в этом втором с половиной поколении и живем, — признается Иомдин.

На сходном уровне находится и «Промт» — самая коммерчески успешная из всех систем машинного перевода, создававшихся в России. Своих успехов «Промт» добивается за счет отказа от тотальности перевода.

— Поймите, — говорит Светлана Соколова, создатель «Промта», — любой перевод любого предложения невозможен в принципе. Если мы хотим, чтобы система работала, мы должны как можно раньше отказаться от понятия «любой». Переводчик всегда будет существовать в ситуации неполного знания, именно этому его и надо учить.

Неполное знание — это проклятая многозначность текста, когда простейшее слово или словосочетание может вдруг выразить чуть-чуть иной смысл, чем тот, что закреплен в словарях. Тут-то машина и садится в лужу. В переносном смысле, конечно.

В свое время в интернете была популярна шутка про перевод с помощью онлайн-версии «Промта» предложения о кошке, родившей трех котят. Фраза «Our cat gave birth to three kittens — two whites and one black» превратилась в «Наш кот родил трех котят — двух белых и одного афроамериканца». Чтобы избавиться от политкорректного котенка, создатели рекомендовали вручную дополнить словарную статью Black, пометив это слово как «неодушевленное».

Когда же будет настоящий переводчик?

— Понимаете, мы живем в пространстве тотальной неоднозначности, — разводит руками Иомдин. — Практически любое высказывание имеет более чем одно значение. Когда

человек пользуется языком, он находится внутри самой жизни и эту неоднозначность очень легко снять. Вот вы приходите домой и говорите: «Я принесла лук». Наверно, ваш муж сразу поймет, что вы принесли: овощ или оружие. Но если взять это высказывание вне жизненного контекста, у нас вообще нет шансов узнать, что оно значит.

Вспомните фразу Бар-Хиллела про детский манеж. С тех пор прошло почти 60 лет, но все машинные переводчики мира уверенно ищут коробки с игрушками в ручках. Подумайте, каким огромным запасом исторических, физических, химических и прочих знаний должна обладать машина, чтобы все понимать про луки, ручки и детские манежи.

И тут проблема машинного перевода предстает как часть куда более широкой темы искусственного интеллекта. Чтобы конкурировать с интеллектом человеческим, ему тоже нужно понимать смыслы.

— В идеале искусственный интеллект — это способность машины создавать самостоятельные суждения, — рассказывает философ, логик и переводчик Делир Лахути. — Если мы имеем тот или иной текст, то машина должна уметь, во-первых, извлечь из него информацию, а во-вторых, знания. Информация — это факты. А знания — это способность выводить из имеющихся фактов неизвестные до сих пор закономерности.

И тут-то перед машинным переводом открываются иные возможности. Чтобы переводить точно, машине не хватает знания контекста — того, что осталось за границами конкретного предложения (как, например, в истории с луком). А теперь представьте, что машинный перевод побратался с другими системами искусственного интеллекта: базами фактических знаний, системами распознавания образов, анализаторами голоса и т. д.

Когда все эти умения сойдутся воедино, вполне вероятно, может получиться машинный переводчик, сопоставимый с человеком. Конечно, машина никогда не будет переводить Шекспира лучше Пастернака. Но там, где важны не художественные параметры, а точность понимания, компьютер теоретически может даже превзойти своего создателя.

Допустим, мы хотим перевести Хемингуэя. Для полноценного понимания смысла, который вложил в свои тексты писатель, нам нужно много чего знать о Гражданской войне в Испании или о быте кубинских рыбаков. При этом интеллектуальные возможности человека-переводчика вполне конечны. Машина же способна оперировать гигабайтами, терабайтами и прочими гигантскими объемами — осталось лишь правильно их связать и создать правила работы.

Что из этого получится? На сайте компании «Промт» размещен шуточный прогноз развития систем машинного перевода. Последний пункт, датированный 2264 годом: «Человек глуп, как мешок опилок, — заявило Устройство 296. — Только абсолютно наивным ученым могло прийти в голову разработать технологию для понимания того, что произносят эти неопрятные куски протоплазмы».

Иллюстрации: *Маша Краснова-Шабаева*