

• Искусственный интеллект: Машинный перевод



Ещё в далёком 1982-м году журнал Popular Science опубликовал список новых перспективных технологий. Эти технологии представляют собой новейшие достижения и инновации начала XXI века в различных областях науки и техники. Несмотря на то, что мнения о степени воздействия на общество, статусе и экономической жизнеспособности некоторых из них существенно различаются, данный список продолжает оставаться одним из самых полных, описывающих те технологии и направления исследований, которые в ближайшее время станут (а большая часть этого списка уже стала) полем для конкурентной борьбы.

Недооценивать коммерческую составляющую этих направлений не стоит. Несомненно, деньги, которые может, к примеру, получить Honda от продажи роботов и деньги, которые выделяются из государственного бюджета – это совсем разные деньги. К примеру, США может себе позволить отдать DARPA 7,6 миллионов для того, чтобы исследовать проблемы регенерации. А что касается компаний крупного и среднего бизнеса, то они вынуждены вкладывать в исследования гораздо меньшие суммы, и только в те области, которые смогут в обозримом будущем окупить эти вложения.

Однако, как показал опыт IBM, можно и частным компаниям получить значительную прибыль для того, чтобы инвестировать её в исследовательские лаборатории.

Так вот, одна из таких перспективных технологий – «Машинный перевод». И мне очень приятно, что у нас в России есть лаборатория из 80 (!) человек, которые всерьез занимаются этой технологией и являются одной из самых сильнейших команд в мире. В прошедший вторник я встретился с директором этой команды, Светланой Соколовой, и взял у неё мини-интервью.

Немного предыстории. В 1991-м году, когда СССР стал разваливаться и политический, экономический и социальный кризисы стали разъедать все государственные учреждения во всех слоях и направлениях, большинство научных учреждений и лабораторий оказались в весьма плачевном состоянии. Очень немногие смогли выжить, большая же часть была расформирована (или сотрудники разошлись сами из-за отсутствия финансирования), эмигрировала за рубеж, или стала «малиновыми пиджаками».



Светлана Соколова, начавшая свою работу в лаборатории инженерной лингвистики Педагогического института имени А.И. Герцена ещё в 1979 году, к 1982-му году стала руководителем группы программистов лаборатории и её старшим научным сотрудником. И, когда грянули «лихие» 90-е, она смогла сохранить команду и создать компанию «ПРОМТ», специализацией которой стал выпуск программ машинного перевода.

Сейчас на российском рынке компания ПРОМТ предлагает линию систем перевода под торговой маркой PROMT для различных групп клиентов: для домашних пользователей, профессиональных переводчиков, а также корпоративных клиентов.

ПРОМТ - обладатель Национальной премии России. Компания также неоднократно признавалась независимыми агентствами наиболее влиятельной компанией в российском компьютерном бизнесе.

23 марта компания собрала в Москве журналистов для того, чтобы объявить о выходе новой, 9-й версии переводчика, а так же рассказать о его особенностях.

По словам Светланы Соколовой, емкость рынка услуг по переводу текстовой информации огромный – более 10 миллиардов долларов в 2010-м году (по данным исследования компанией Common Sense Advisory). Из них 2 млрд. долларов – услуги машинного перевода (Winter Green Research).

Светлана, расскажите, для каких задач используется машинный перевод?

Машинный перевод чаще всего используется для общения, для получения информации из текстов на иностранном языке и для получения перевода текста с целью публикации. Когда мы только начинали свои разработки в начале 80-х, ещё никто не предполагал, что будет так развит интернет и будет такая массовая потребность в обмене информацией. Поэтому сейчас очень большой объём машинного перевода применяется для общения в чатах, социальных сетях и форумах.

Вторая цель – для получения информации из текстов на иностранном языке более очевидна, и, собственно, она как раз и ставилась разработчиками изначально. Что же касается третьей - получения перевода текста с целью публикации, то она так же явилась трендом последних лет, в связи с тем, что количество СМИ и новостных онлайн-блогов увеличилось в разы.

Машинный перевод и искусственный интеллект - насколько они связаны?

Вообще, понятие ИИ родилось практически одновременно с задачей автоматической обработки текстов. Сейчас наша команда исследований и разработки составляет 80 человек. Из них более 60 занимаются лингвистическими исследованиями и развитием ядра системы – модулей перевода. Так же есть специальная выделенная группа, которая занимается разработкой средств разработки. Дело в том, что таких средств, которые нам необходимы, их просто нет в природе и мы сами должны их использовать. Что это за средства?

Во-первых, это специальные тестеры для поддержки эволюционирования системы. С их помощью мы можем на больших

массивах текстовой информации следить за изменением качества перевода, построении обработки конструкций и т.д.

Во-вторых, это автоматические генераторы объектов тестирования – компонентов лингвистической базы данных, структуры, которые анализируют программы и т.д. Эти генераторы так же ускоряют процесс разработки.

В-третьих у нас есть специальные инструменты сравнения парсингов, которые позволяют непротиворечиво развивать алгоритмы синтаксического и семантического анализа.

В-четвёртых, средства для поиска в текстовых корпусах любого объёма. Мы обрабатываем колоссальные объёмы текстовых корпусов, из которых мы извлекаем правила и семантические связи, которые включаем в нашу систему. И для этой обработки нужны специальные средства поиска, так же разработанные нами.

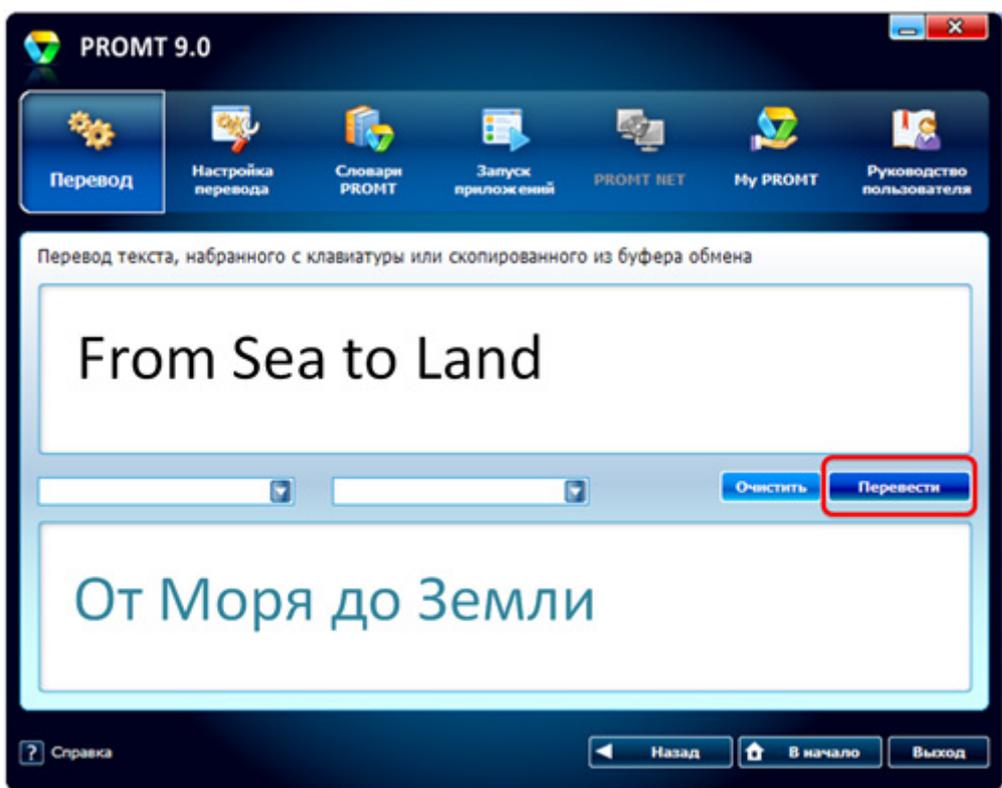
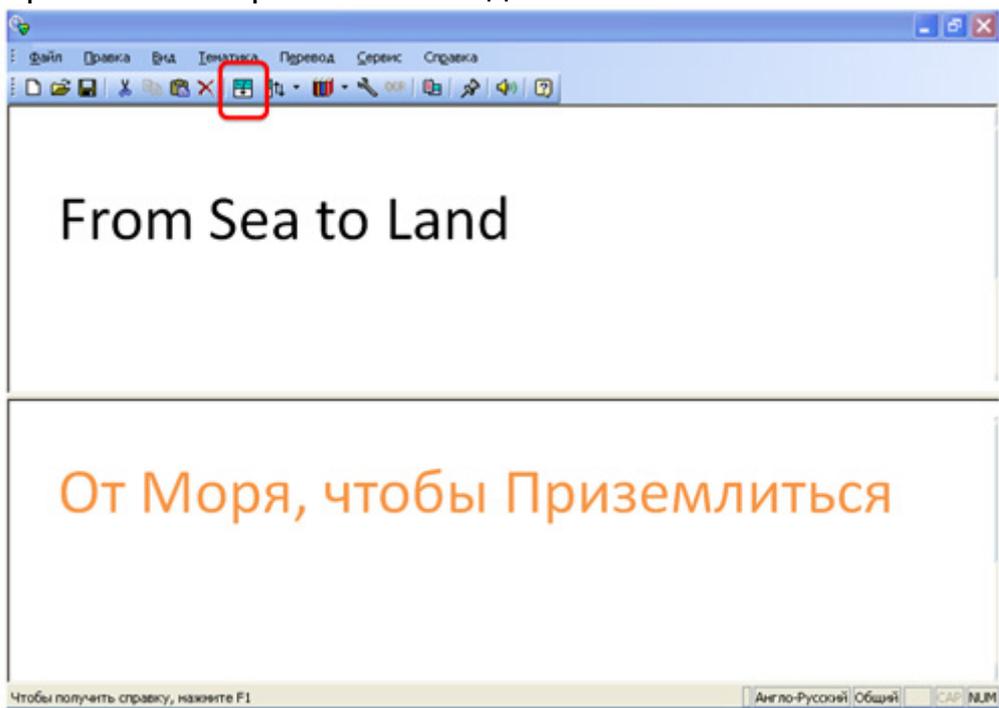
И в-пятых – статистические модели языка. Средство, которое позволяет создавать правила на основе статистических методов.

За последние два года мы смогли научить нашу систему общаться со сложно устроенными семантическими моделями и заложили технологическую основу для развития семантического компонента системы. Кроме того, мы эффективно применили в разработке корпусные и статистические методы для значительного расширения синтаксических и семантических правил.

Следующая версия нашей системы будет гибридной – объединит rule-based систему со статистической, что позволит повысить качество перевода.

Вообще, если говорить об искусственном интеллекте, то, применительно к машинному переводу, чем большего качества мы стремимся добиться, тем больше сил, финансовых и человеческих ресурсов в это приходится вкладывать. Поверьте, не каждая лаборатория в мире по проблемам ИИ сможет похвастаться штатом в 80 человек – лингвистов, математиков и аналитиков. Но мы понимаем, что для создания принципиально нового поколения систем машинного перевода нужен штат в несколько раз больше.

Сравнение старого и нового движка:



Заключение

Для того, чтобы ИИ смог сравниться с человеческим, человечеству придётся решить ещё массу проблем. Трудно вообразить то количество направлений и навыков, которые такими простыми и лёгкими кажутся для нас (к примеру, распознать картинку CAPTCHA) и пока ещё

непреодолимы для роботов. И тут, так же, как и в разработках по проекту SENS, ключевую роль играет финансирование.



По словам **Ирины Панасенковой**, директора по продажам компании ПРОМТ, дела у компании идут неплохо, даже несмотря на кризис (а в некоторых сегментах даже благодаря ему, ведь ПРОМТ - возможность снизить затраты на переводчиков). «Наш новый принцип – создавать уникальные продукты для четко выделенных сегментов. Новые решения PROMT Language Service Provider 9.0 и PROMT Freelance 9.0 обеспечивают профессиональную поддержку переводческих проектов как в рамках корпоративной сети, так и с участием переводчиков, работающих вне офиса. Своевременность выхода новой версии программ PROMT подтверждается общей ситуацией на мировом рынке. По данным исследований международной ассоциации TAUS, 81,8% компаний планируют в ближайшие два года начать использовать решения для машинного перевода для достижения таких целей, как сокращение издержек и повышение точности перевода (69,7%), расширение бизнеса (35,5%), организация поддержки клиентов из разных стран в реальном времени (37,4%). При этом в качестве результата от внедрения систем машинного перевода корпоративные клиенты ожидают снижение стоимости переводимого контента (58,1%), переход к инновационным методам ведения бизнеса (36,0%), автоматизацию бизнес-процессов (18,5%)» - говорит она.

Кроме того, Ирина отмечает, что в течение ближайшего года будет запущен проект SAAS (сперва на европейском рынке, который более готов к этой бизнес-модели, а потом и в России), так что финансовое обеспечение исследовательской лаборатории будет только расти.

Однако, другие лаборатории, специализирующиеся на развитии ИИ не могут похвастаться такими бюджетами.

К примеру, очень близко к машинному переводу находятся разработчики ИИ, способного распознать слитную речь. Но, к сожалению, рынок услуг распознавания речи пока значительно меньше рынка МП, соответственно там и инвестиций меньше. Если продукты ПРОМТ используют все – начиная от госсектора, заканчивая частными лицами, то распознавания речи – это очень узкий сегмент – коллцентры, системы прослушивания, телекоммуникационные приложения...

Какая из лабораторий первой порвёт ленточку, покажет время. Пока же ИИ ещё далеко до человеческих способностей.

Оригинал статьи – на <http://trendclub.ru/3975>.