К.В. Ламин

В последнее время становится всё больше и больше статей, посвящённых системам распознавания речи, в том числе и в изданиях, редко пишущих о науке и технике. Однако, читая такие статьи, невольно оказываешься в состоянии очевидца, читающего репортаж о событии в вечерней газете. Кто в такой ситуации был, хорошо помнит ощущение несусветной чуши и откровенного вранья, с мелкими вкраплениями вроде бы правильных фактов. При этом путание миллионов с миллиардами, замена всех букв в фамилиях еще, можно сказать, мелкие шалости.

Такие статьи, по-сути, — это попытки неспециалистов бросить свой обыденный взгляд на довольно замысловатые вещи и поделиться этим сокровищем мысли с другими неспециалистами. Этот интерес к теме сам по себе не плох, но пора, наконец, и специалистам потратить часть своего драгоценного времени на объяснение того, чем они, собственно, занимаются. Это придётся сделать ещё и потому, что решения о государственном финансировании разработок делаются частенько на основе неверных представлений о сложности и важности задач, сформированных такими поверхностными статьями. Ожидания и задачи промышленности в использовании СРР также искажаются, но в меньшей степени и по вполне понятным причинам. Я призываю к дискуссии на затрагиваемые в этой статье темы и, разумеется, ожидаю уточнений и дополнений. Очень хотелось бы услышать голос тех, кто использует или хотел бы использовать СРР в бизнесе. Иначе, неизбежны искажения и натяжки, свойственные специалистам и энтузиастам.

Итак, что волнует публику? Публику в Одессе и в других городах волнует вопрос: «что мы с этого будем иметь?». Попробуем дать на это ответ. При этом, как опытный психоаналитик, оставим на время то, что по этому поводу думает сама публика. Заглянем, так сказать, поглубже. На поверхностном уровне у публики есть стереотипы восприятия всех составляющих искусственного интеллекта, сформированные Азимовым, Терминатором и ещё незнамо кем. В этом образе СРР — это часть искусственного гомункулуса, слеплённого с неясными целями. Этот стереотип настолько устойчив, что используется хитрыми PR-менеджерами для продвижения той или иной компании. Вспомните щенячий восторг телевизионщиков вокруг разного рода «человекоподобных» роботов! Следуя этой привнесённой извне концепции, несчастный обыватель воспринимает СРР исключительно в виде полностью разумного, но желательно глуповатого собеседника на отвлеченные темы. Из этого нехитрого образчика коллективного бессознательного и происходит задача свободной диктовки текстов. Спроси любого — «зачем нужны СРР?» и получишь — «Чтоб текст в компьютер диктовать!».

По этому бесславному пути пошли многие команды разработчиков и благополучно сгинули. И ещё пойдут и ещё сгинут, унося в небытие миллионные инвестиции. Унесли, между тем, уже больше полумиллиарда долларов, по самым скромным подсчётам. Рискну оказаться непопулярным в среде разработчиков, потребляющих эти инвестиции, но скажу: задача диктовки свободного текста — это не задача распознавания вообще, более того, это почти никому не нужно! В самом деле, любому специалисту ясно, что человеческая речь — это средство коммуникации хоть сколько-нибудь разумных существ. Добрая треть информации воспринимается на слух исключительно благодаря тому, что мы понимаем, о чём идет речь. Американские морпехи, к примеру, вместо систем шифрования использовали в радиопереговорах индейцев племени Навахо, передающих сообщения на своем языке, справедливо полагая, что даже записать транскрипцию противник будет не в состоянии. По моему глубокому убеждению, записать на слух можно только известные тебе понятия, иначе придётся прибегнуть к диктовке по буквам. Имея такую фамилию, могу говорить это с полной научной достоверностью. Непривычная для русского уха, на слух она записывается неверно в 90% случаев!

Не хочу сказать, что такая система невозможна в принципе, более того, при условии ограниченного набора понятий она вполне реализуема, однако это не задача распознавания и чисто акустическими и лингвистическими методами вряд ли можно добиться больше, чем диктовка SMS.

Теперь немного о том, где найдёт применение такая система. Для этого разберём для начала, где нам совершенно необходим печатный текст. Это издательское дело и юриспруденция. Точка. … Обескураживает, не так ли? Во всех остальных областях текст используется исключительно из-за компактности при записи, хранении и передаче. И из-за привычки, конечно. 99% информации, которой мы обмениваемся, не стоит того, чтобы переводить её в текст. Если бы у древних были бы сегодняшние способы сохранения, передачи и обработки голосовой информации то скорее всего письменность бы так и не изобрели. Забавно, что задача ввода текста в компьютер возникла ещё и потому, что всем печатать или писать собственноручно попросту лень. Не зря же те, кто мог себе это позволить, диктовали свои перлы машинисткам. Т.е. представление речи в виде текста — явление прежде всего экономического плана. Однако мы живем в эпоху, когда экономические стимулы к этому постепенно исчезают. Особо бедные слои населения ещё используют SMS из-за того, что это дешевле голосовых сервисов, но это отмирающая технология. Вся электронная переписка, переговоры, СМИ прекрасно поддаются переводу в аудио-видеоформат. Давно ли Вы писали письмо на бумаге? Не загнало ли газеты в узкие ниши телевидение? Не исчезнут ли газеты вообще с приходом сотен и тысяч каналов IP-TV? Пока что работать с текстом всё же удобнее, чем с голосовыми файлами, главным образом потому, что не решена с приемлемым качеством задача быстрого поиска фрагментов и редактирования записей. Но это пока! Так стоит ли заниматься задачей распознавания свободного текста? Стоит конечно, но понимая, что заниматься нужно созданием систем «понимания» свободного текста, т.е. искусственной семантической сетью с человеческим масштабом количества узлов, упираясь при этом в совершенную недостаточность современных систем представления данных и отсутствие приемлемой модели параллельных вычислений. Учитывая, что диктовка нужна только писателям и юристам, которые вряд ли окупят вложения ввиду своего малого количества. Какой из инвесторов вложит свои деньги в такой проект? А чужие или государственные — это, конечно, каждый сможет! Стоит ли приветствовать такие вложения? В условиях конкуренции за ресурсы — вряд ли! Т. е. это задача для сверхбогатого университета. С учебными целями и без надежды на результат.

Итак, мы натолкнулись на то, что свободный текст задиктовать в компьютер неимоверно сложно прежде всего потому, что компьютер не понимает диктуемое. На сегодня это

так. Так что же, про использование СРР в сегодняшней жизни можно забыть? Вовсе нет! Существует огромный класс задач для СРР. Это человеко-машнные интерфейсы. Эта отрасль знания настолько пахнет нафталином, что любые улучшения пойдут на пользу. Сразу понятно, что всякая свободная диктовка здесь не нужна. Мир устройства, с которым мы вознамеримся поговорить, ограничен его функциональным назначением. Грубо говоря, нам нужно, чтобы устройство распознавало те команды, на которые оно в принципе способно отреагировать в текущий момент времени. Более того, не очень необходима, хотя и желательна более-менее свободная манера выдачи команд. В самом деле, если при помощи команд типа «Стой, ать-два!» можно победить в войне, то стоит ли большего требовать от автоматических устройств. Таким образом, приняв вышеуказанные ограничения вполне возможно свести словарь распознаваемых фраз к достаточно разумному числу. Что это даёт? Даёт реальную работоспособность таких интерфейсов. При таких самоограничениях возможны коммерческие интерфейсы к большинству устройств и программ. Что мешает повсеместному внедрению таких систем? Инерция производителей и потребителей. Она вполне понятна, сегодняшние «интуитивные» интерфейсы ужасны, но еще больший ужас вызывает у потребителей необходимость переучиваться другому способу отдачи команд. Отвращение производителей к столь глубоким новациям, наверное, ещё больше. Но быстро или медленно, этот процесс будет идти, завоевывая всё новых приверженцев. Быстро он пойдет в том случае, если мы научим компьютеры распознавать и разумно реагировать на ограниченный по смыслу набор команд, высказанный допустимыми в естественном языке способами, ограничивая лишь использование слэнга и особой витиеватости. В таком случае переучиваться не придётся, ведь разговаривать мы уже худо-бедно научились в детстве.

Какие требования к интерфейсным системам предъявляются по минимуму и какие улучшения этого минимума выльются в конкурентные преимущества?

Предположим, мы вознамерились дать рядовым разработчикам интерфейсов некую библиотеку подпрограмм, с тем, чтобы они могли, не обладая специальными знаниями, внедрять такие интерфейсы в свои разработки.

Относительно общего словаря системы. Очевидно, что он должен быть максимально большим и включать в себя возможность добавления новых и редких слов. При этом соображения о том, что, дескать, человек оперирует всего лишь 10–12 тысячами слов, здесь не применимы, т.к. нам нужно универсальное решение как в плане задач, так и в плане пользователей. Таким образом 80–100 тыс. слов — вполне разумные цифры.

Динамический словарь — нужен для того самого самоограничения, связанного с учётом ограниченного числа возможных действий объекта управления в тот или иной момент времени. Наличие таких самоограничений благотворно сказывается на точности распознавания. Понятно, что требования к такой подсистеме выражаются в том, что она вообще должна быть и при этом удобна в использовании.

Распознавание отдельных слов. Речь идёт о заполнении полей в базах данных и различных запросах. Очевидно, что для целей дальнейшей обработки

количество альтернатив в одном поле (т.е. в текущем состоянии динамического словаря) редко превышает несколько десятков. Исключение составляют универсальные поисковые системы, для которых динамический словарь расширяется до размера общего словаря системы. Система должна распознавать и короткие слова достаточно хорошо.

Распознавание слитных фраз — совершенно необходимое свойство системы. В первую очередь из-за того, что люди привыкли говорить слитно.

Количество одновременно имеющих смысл командных фраз различно для разных объектов управления. Но для большинства приложений достаточно возможности распознавать одновременно несколько тысяч командных фраз. Т.е. чем больше это количество, тем удобнее интерфейс и тем большее количество объектов может управляться таким образом. Чем больше размер динамического словаря — тем шире рынок. При этом если мы построим кривую удобства в зависимости от размера динамического словаря, то в области малых значений (меньше 20) она пойдет вверх, а затем долгое время будет снижаться или стоять на месте, и лишь в области более тысячи начнёт повышаться и станет выше, чем в области малых значения. Я бы назвал это «долиной лени и забывчивости». Дело в том, что единичные команды довольно легко запоминаются, и поэтому достаточно запомнить «сим-сим откройся» — и голосовое управление дверным механизмом освоено! Нужно как минимум крепко выпить, чтобы забыть несколько команд. А вот когда их становится больше — запоминать всё труднее и приходится тратить время и усилия на обучение интерфейсу. Когда мы можем распознать тысячи команд, а объект управления имеет только десятки функций, вполне возможно распознавать все возможные синонимичные варианты команд, выполняя несложные ограничения типа «иностранными словами не выражаться!». При этом нужда в запоминании команд отпадает, надо лишь знать функциональные возможности объекта. Сильно в этом помогает использование диалогов с понятными и чёткими указаниями пользователю, какого рода информацию от него ожидают. При этом удается разбить динамический словарь на ряд состояний с приемлемым размером в каждом из них. Это, кстати говоря, вполне человеческая манера.

Количество неверно распознанных команд. Для не шибко динамичных и не самых ответственных объектов управления ограничения накладывают лишь соображения комфорта. Однако они достаточно жёсткие — менее 2–3 ошибок на сотню команд. В самом деле, нажимать на курок с помощью речи вряд ли стоит, а в большинстве случаев достаточно, чтобы в случае ошибки или ложного срабатывания хорошо распознавалась команда «назад!». При этом стоит обращать внимание именно на количество ошибок в сотне команд, т.к. «процент распознавания слов» и т.п. не совсем корректные показатели. Существует ещё ряд показателей качества СРР именно в области безошибочности, и для точной характеристики системы одного показателя мало, но всё же предлагаемый показатель достаточен для коммерческой оценки СРР.

Минимально необходимое соотношение сигнал/шум для получения нужного показателя безошибочности. Устойчивость к искажениям сигнала. Эти показатели связаны напрямую со стоимостью необходимых систем ввода и оцифровки звука, а также с возможным окружающим объект шумом. Опасайтесь прекрасных данных системы, снятых в особо тихих помещениях с помощью системы микрофонов стоимостью в несколько сот тысяч долларов! Скорее всего, распознавание в такой системе построено на тонких нюансах, которые пропадут при реальных условиях. Т.е. для коммерческой системы негоже полагаться на системы шумоочистки и шумоподавления, СРР должна быть робастна по шумам.

Быстродействие — коммерческий показатель. Очевидно, что время срабатывания для интерфейсов не более 1,5 секунд. При этом на коммерческие перспективы влияет минимальная требуемая вычислительная производительность, при которой выполняется условие про 1,5 секунды. Не стоит думать, что вам кто-нибудь из коммерческих заказчиков отдаст все вычислительные ресурсы только на интерфейс. Если 10% загрузки процессора выклянчите — считайте повезло. Сможете уложиться в ресурсы встроенных систем (КПК, бортовые RISC-процессоры) — для вас откроются дополнительные рынки.

Устойчивость результатов с разными дикторами, акцентами и желательно с разными языками. С точки зрения удобства использования система должна работать без предварительной настройки на диктора, иногда допустима тонкая подстройка в процессе работы, незаметная для пользователя. Требование к многоязычности очевидно происходит из глобального характера современной экономики. Поэтому часто даже весьма несовершенная, но многоязычная система побеждает в тендерах глобальных корпораций.

Что касается превышения этих требований, то всё зависит от количества ниш, которые вы хотите занять. При выполнении минимальных требований ниша на рынке обязательно найдётся! Из животрепещущих тем я бы назвал увеличение размера динамического словаря, многоязычность и достижение большей устойчивости к шумам и искажениям. В остальных областях возможности лучших современных коммерческих СРР вполне приемлемы. Для нелучших — задача номер один достичь минимальной планки и найти своё «место под солнцем».

И перестаньте, наконец, заниматься системами диктовки!

Ламин Константин,

CEO Speereo Software UK
Председатель Совета директоров
ЗАО «Титан-Информационный Сервис»