Непрочитанный доклад
Александра Семёновича Нариньяни

Голубые города

26 апреля 2010 года ушел из жизни Александр Семенович Нариньяни. Его смерть всколыхнула воспоминания о более чем полувековом знакомстве, профессиональном сотрудничестве и дружбе с этим неординарным, замечательным ученым и человеком.

Так случилось, что в 1955 году мы оказались с Александром Семеновичем (далее с Сашей) в одном доме и в одном подъезде, ездили в одном и том же лифте. Тогда никто из нас, видимо, не подозревал, что нам — казалось, весьма далеким по будущей профессии — предстоит вступить в весьма близкий научный контакт. Я учился на филологическом факультете МГУ на отделении классической филологии, а он — в МИФИ, сугубо техническом вузе. Однако каждый из нас проделал довольно существенный, непрямолинейный путь к занятиям, связанным с той областью, которая ныне называется компьютерная лингвистика.

Саша провел свою дипломную практику в новосибирском Академгородке, а после окончания МИФИ в 1962 году резко изменил место своего проживания и отправился создавать и обживать Академгородок, суливший новые горизонты для академической науки. Там он начинает в 1962 году свою академическую карьеру в должности старшего лаборанта в Институте математики, после открытия ВЦ переходит туда.

Академгородок славился своим вольнодумством и активной культурной жизнью. Там действовал клуб «Под интегралом», где выступали те артисты и барды, которым был закрыт доступ к зрителям и поклонникам в столицах, организовывались небывалые художественные выставки, проводились диспуты в условиях небывалой демократии. И там кипела активная научная жизнь. Возможно, под влиянием своего научного руководителя академика Андрея Петровича Ершова, усиленно занимавшегося информатикой и информатизацией, Саша обратился к лингвистике как науке, связанной с информационными и интеллектуальными технологиями. Это потребовало от него довольно существенного погружения в лингвистику.

Я вновь встретился с ним в 70-е годы, но теперь — как с вполне владеющим лингвистической терминологией, проблематикой и менталитетом. Это было время инициированного им (теперь уже более тридцати лет назад) проекта междисциплинарного семинара, объединявшего лингвистов, инженеров, специалистов по искусственному интеллекту, психологов и даже философов. Этот уникальный ежегодный зимний семинар проходил в 1975–1989 гг. в Эстонии, особенно при участии коллег из Тартуского университета. Постепенно эти собрания приобретали черты тематических семинаров — «Модели общения», «Диалог». Последнее название сохранилось и по сей день.

Саша был генератором глобальных проблем и успешным продюсером больших проектов. В течение более десяти лет он сотрудничал с нашей университетской группой в форме тогдашних грантов, по тем временам — хоздоговоров. Не имея никаких собственных финансовых ресурсов, он ухитрялся доставать их у тогдашних меценатов — различных министерств, которым советский Госплан выделял фонды на науку. Для меня вся эта его деятельность была полной загадкой, мы имели дело только с ним и отчитывались только перед ним. Формат хоздоговоров позволял создавать ставки и расширять кадровый состав группы.

В 70–80-е годы мы очень сблизились с Сашей. Мы много спорили на наших workshop'ах и мозговых — как он любил говорить — штурмах, но наши споры не были никогда омрачены отстаиванием личных амбиций. Главной целью был добрый поиск правильных решений.

90-е годы были для Саши трудным испытанием. Еще в 1989 году он оставил Академгородок, мало походивший на тот, с которого все начиналось. Я знаю об этом понаслышке и понимаю только общую траекторию изменений. Саша вернулся в Москву, увлеченный идеей создания Голубого города, где были бы воплощены усохшие идеалы Академгородка. Он чуть ли не подыскал в Подмосковье место для этого Города и сколачивал команду единомышленников. Но стране было уже не до Голубых городов. Когда рухнула старая система, все привычные рычаги влияния исчезли. Саша создал в 1999 году ведомственный Институт искусственного интеллекта, способ существования которого был мне непонятен, но в новую ситуацию вписаться было трудно. Он мыслил широкомасштабными проектами на десятилетия, а искать и находить средства к сиюминутному проживанию не было сильной стороной его интеллекта. Мало кто знает, что последние 10 лет, продолжая работать, он фактически не получал зарплаты.

Все мы сейчас еще не готовы осмыслить всё, что было сделано Сашей. Но мне кажется, что главная Сашина черта — романтизм: вера во всесилие добра и доброго знания, в возможность делать благое дело безо всякой корысти и пересчета на возможные личные дивиденты. Он, по существу, жил в неком виртуальном мире, где все устроено просто — заботиться не о себе, а об идее. В этом была его сила и его слабость.

В последние годы чувствовалось, что он угнетен многими несбывшимися ожиданиями. Однако его постоянно пытливая мысль, направленная на идеальную цель с далеко идущими последствиями продолжала напряженную работу. Публикуемая в материалах конференции его последняя статья оказалась как бы его завещанием. В этой статье утверждается идея государственного масштаба. Нынешнее состояние науки пока еще дает возможность России вытянуть счастливый билет и сделать прорывный скачок в наиболее продвинутой у нас в стране области знания. Именно этот билет есть естественная и доступная для России сфера получения довольно быстрого роста в экономике и культуре, требующая инвестиций не в нефтепроводы и даже не в нанотехнологии, а в человеческие мозги и научные знания.

Саша говорит власти: «Используйте реальный шанс, если вас заботит будущее России как процветающего современного государства. А я ничего не могу сделать больше, чем указать на эту возможность, моя миссия выполнена, я могу уходить».

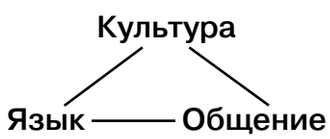
А. Е. Кибрик
3 мая 2010

Русский язык как национальная программа

Нариньяни А. С.

1. Язык как фундамент культуры

Язык — необходимый компонент базового треугольника любой мировой культуры, нет прецедентов, когда бы язык разрушался, а большая цивилизация оставалась.



Однако меняются этапы развития человечества и формирующие их технологии-доминанты. Сегодня будущее любой культуры определяется уровнем ее симбиоза с информационно-коммуникативными технологиями (ИКТ), и прежде всего погружением ее языка в глобальное Интернет-пространство, из которого на наших глазах формируется электронная «нервная система» современного человечества.

Если до сих пор та или иная страна определялось своей географической территорией и ролью в истории, то теперь борьба перешла во всеобщую информационную сеть, где развертывается следующий этап соревнования за место в глобальном мироустройстве.

Сегодня обработка информации на естественном языке является основой практически любого вида деятельности и поэтому ее роль особенно важна для подавляющего большинства приложений, тем более что в ближайшей перспективе широкое внедрение текстовых и речевых технологий сделает взаимодействие с компьютером по-настоящему массовым и естественным.

Для России компьютерные технологии обработки языковой информации (КТОЯ) важны еще и для сохранения быстро сокращающейся виртуальной русскоязычной территории, определяющей ее связи с многомиллионной диаспорой за рубежом, а также для поддержки тех международных коммуникационных контактов, которые сложились за вторую половину двадцатого века.

КТОЯ уже пришли в государственную и общественную сферы, становятся частью дома и личной жизни. Таким образом, обработка языковой информации и по важности и по масштабу – проблема национального уровня. Однако именно ее масштабность ставит под вопрос ее посильность для основного большинства языков и стран. Лишним доказательством этого является тот факт, что пока только небольшое число языков успешны в некоторых секторах приложений. Причем даже эти результаты можно оценить как достаточно умеренные, несмотря на огромный объем вложенных здесь усилий и средств.

2. Национальная программа

2.1. В известной степени обсуждаемая ситуация сходна с положением в атомной физике на границе 30х и 40х годов, когда перспектива появления атомного оружия потребовала крайней мобилизации ресурсов — прежде всего, интеллектуальных — для форсированного перехода от *исследовательской науки* к дорогой и сложной области оборонной *индустрии*.

Можно упрекнуть меня в чрезмерной драматизации. Однако сегодня, как и тогда, решается ключевой вопрос судьбы страны, определяется будущее нашей культуры в ряду основных мировых культур. И сегодня снова необходим рывок, обеспечивающий переход от науки, привыкшей работать в режиме НИР, в национальную отрасль, способную не только масштабно решать прикладные задачи, но активно работать на экспорт.

В таком прорыве мы можем опираться: на отечественные школы лингвистики и информатики, которые остаются пока одними из лучших в мире. Реализация этого потенциала должна позволить создать те инновационные языковые технологии, которые обеспечат необходимый успех в борьбе за ведущее место России в глобальном распределении ролей.

2.2. Очевидно, что координация усилий такого масштаба возможна только на уровне Национальной Программы, обеспечивающей самую продуктивную кооперацию участвующих в ней научных, прикладных и коммерческих сил.

Такая Программа (обозначим ее «Русский Язык-ИКТ») должна совместить четыре ключевые составляющие:

- ⇨ *фундаментальные работы по интеграции лингвистических ресурсов русского языка;*
- ⇨ *комплекс прорывных программных КТОЯ, в первую очередь, для русского языка;*
- ⇨ *механизмы маркетинга создаваемых продуктов на отечественном и мировом рынке;*
- ⇨ *подготовку высококвалифицированных кадров и обучение пользователей.*

В стране сегодня работает несколько программ поддержки работ в области русского языка, однако они ограничены финансово, не координированы и ориентированы на достаточно специальные сектора. Эти работы могут служить дополнением к предлагаемой Программе или быть ее частью, но никак не ее альтернативой.

Обеспечить ядро Программы может только консорциум фирм и научных коллективов, способный при участии инвесторов и поддержке государства обеспечить ее реализацию — создание комплекса интеллектуальных ИКТ нового поколения, ориентированных на эффективную обработку информации на русском и, возможно, ряде других языков.

2.3. Значение Программы для общего фронта работ, определяющих сохранение нашей страной быстро утрачиваемого ею статуса научной сверхдержавы, переоценить нельзя, поскольку эффективность КТОЯ и базирующихся на них современных каналов обмена информацией является не менее важной составляющей прогресса, чем уровень кадров и современное оборудование.

Актуальность и стратегическая важность проблемы подтверждается широким фронтом работ в данной области и объемом средств, которые тратятся на них во всех ведущих странах. Однако, текущий уровень основной массы таких продуктов не обеспечивает пока результатов, гарантирующих их прикладную эффективность и их массовое внедрение.

Потенциальные участники такой Программы имеют многолетний опыт в этой области, подтверждающий как высокий уровень разрабатываемых ими технологий, так и оригинальность положенных в их основу ноу-хау. Именно это сочетание опыта, принципиально новых подходов и сложившихся высококвалифицированных команд и, главное, имеющихся заделов, обеспечивают Программе высокую вероятность прорыва в создании качественной технологии следующего поколения.

В судьбе страны Программа «Русский Язык-ИКТ» должна выполнить не менее ключевую функцию, чем проекты, включаемые руководством страны в спектр стратегических инновационных направлений. Если в развитии нано- и био-технологий Россия может сотрудничать со всем миром, то место русского языка в ИКТ — это будущее *нашей страны* и ее культуры — определяется только здесь совокупностью именно наших усилий.

Потеря времени будет усложнять эту задачу на порядки, а каждый выигранный год даст прикладной и моральный выигрыш, намного превышающий объем инвестиций, требуемый для этого необходимого рывка сегодня.

3. Структура Программы

3.1. Структура проблемной области КТОЯ в самом общем виде может быть представлена следующим образом:

1. *Содержание и понимание (смысла текста, знания об области приложения, онтологии)*
2. *Анализ текста*
   - 2.1. Индексация и классификация текстов
   - 2.2. Извлечение содержания текста
   - 2.3. Понимание сообщения конкретной тематики
   - 2.4. Понимание текста пользователя в контексте диалога
   - 2.5. Определение синтаксиса и стиля и их корректировка
   - 2.6. OCR
3. *Генерация текста*
   - 3.1. Вопрос, Сообщение, Команда, Ответ в контексте диалога
   - 3.2. Документ конкретной тематики и жанра
4. *Перевод (автоматизированный и автоматический)*
5. *Поисковые системы*
   - 5.1. Файловые системы
   - 5.2. Однородные текстовые массивы, БД, электронные архивы и библиотеки
   - 5.3. Плохо структурированные массивы
   - 5.4. Локальные сети
   - 5.5. Интернет
6. *Диалог «человек — компьютер»*
7. *Лексика*
   - 7.1. Словари одноязычные, двуязычные, многоязычные

7.2. Корпусные БД
7.3. Тезаурус
8. *Гипертексты*
9. *Стандарты*

3.2. Естественно, что такая общая классификация не может претендовать на полноту, поскольку должна покрывать весь спектр проблематики от теории до коммерческих продуктов.

Фундамент Программы образуют три основных взаимосвязанных группы работ, отражающие следующие направления тематики КТОЯ:

– Формальные описания основных составляющих знаний о языке,
– Систематизированная лексика (словари),
– Аппарат представления знаний и его приложения в КТОЯ.

Каждая из этих групп рассмотрена ниже в соответствующем подразделе.

3.3. Формализация лингвистических знаний о языке — необходимый компонент создания КТОЯ, основные составляющие лингвистического обеспечения которого включаю такие разделы как:

- *Морфология*
- *Лексическая семантика*
- *Синтаксис*
- *Структура текста*
- *Структура документа*
- *Организация диалога.*

Каждая из этих составляющих складывается из:

⇨ разработки формальных средств описания,
⇨ самого описания соответствующей части лингвистических знаний,
⇨ программных средств использования этих знаний в КТОЯ.

Для первой половины этих — наиболее проработанных и широко используемых — составляющих сегодня особо важными являются лингвистические НИОКР, направленные на доработку полного их описания, поскольку именно оно определяет достаточность и эффективность соответствующих компонентов КТОЯ. При этом обеспечение соответствующих формальных и программных средств может рассматриваться как задача относительно техническая. Для второй половины списка дополнительных усилий требует разработка всех трех компонентов.

3.4. Лексика — совокупность слов и устойчивых словосочетаний того или иного языка, определенной сферы деятельности или конкретной совокупности текстов, такая же ключевая составляющая КТОЯ, как и остальные две. Основные компоненты данного раздела обеспечивают несколько уровней работ от накопления языкового материала (первые два раздела) до формирования словарей самого верхнего уровня (тезаурусы), организующих лексику по семантическому принципу с отражением базовых отношений. Данные компоненты включают:

- *Корпус русского языка со всеми его более специальными составляющими,*
- *Проблемная лексика (каталоги, справочники, терминологические словари и т. п.),*
- *Словари одноязычные, включая базовый интегральный словарь русского языка,*
- *Словари двуязычные / многоязычные,*
- *Тезаурусы.*

Эта подгруппа проектов включает как содержательное наполнение перечисленных составляющих лексического фонда КТОЯ, так и разработку обслуживающих их технологий. Таким образом, как и в предыдущей подгруппе, каждая из этих составляющих складывается из:

⇨ разработки формальных методов для соответствующих групп лексики,
⇨ содержательного наполнения,
⇨ программных средств оформления лексики в компоненты КТОЯ.

Масштаб работы и распределение весов указанных трех частей для перечисленных составляющих различаются не менее чем на порядок и зависят от:

- степени их проработанности,
- полноты и объема содержательного наполнения,
- необходимости каждой из них для тех или иных приложений.

Практически все перечисленные компоненты являются необходимыми и для речевых технологий, особенно для ограниченных предметных областей, где необходимо учитывать предметную направленность текста при разрешении неоднозначности распознавания.

3.5. Наиболее эффективные КТОЯ связаны с использованием смысла ЕЯ сообщения, который в свою очередь неотъемлем от контекста этого сообщения, т. е. знаний о *прагматике дискурса* и о *предметной области*, к которым оно имеет отношение. Поэтому третьей необходимой составляющей КТОЯ является как сам аппарат представления знаний, так и его применение в соответствующих областях. Сюда относятся:

- *Технология представления знаний,*
- *Онтологии,*
- *Модели предметной области,*
- *Представление смысла текста.*

Естественно, что технологии извлечения смысла текстового сообщения включают все эти компоненты, которые должны войти как необходимые направления в состав Программы.

4. Базовые составляющие Программы

4.1. Таким образом, комплекс прорывных КТОЯ должен охватить все основные составляющие языковых

ИКТ следующего поколения, включающие такие ключевые области приложений как:

- *Содержательная обработка потоков текстовых сообщений,*
- *Эффективный интеллектуальный поиск,*
- *Качественный документооборот,*
- *Обработка текстовых данных в системах поддержки принятия решений,*
- *Электронные архивы и библиотеки нового поколения,*
- *Значительное повышение качества машинного перевода,*
- *Интеллектуальные системы в образовании и медицине,*
- *Комплекс речевых технологий,*
- *Диалоговые языковые интерфейсы к прикладным системам,*
- *Понимание текстов в ограниченной предметной области,*
- *Системы двойного назначения и др.*

Понятно, что для успеха Программы она должна опираться на такие базовые составляющие как:

→ Компьютерный словарный фонд русского языка,
→ Качественные технологии анализа и синтеза текстовой и речевой информации,
→ Технологии понимания текста.

Ниже две составляющие будут рассмотрены более детально.

4.2. *Компьютерный словарный фонд русского языка.* Здесь в основу Программы входят три базовых проекта:

• «Корпус» — работы по созданию корпуса русского языка ведутся коллективом энтузиастов (текущий объём более 140 млн. слов) при минимальном финансировании. Программа требует доведения сделанного до готовности по объёму и качеству, а также по возможности дальнейшего развития и пополнения.

• «Словарь» — интегральный словарь русского языка. Сегодня электронных словарей русского языка разного качества многие десятки, но все они далеки от полноты и завершенности. Для приложений и для самой лингвистики необходим единый комплексный «государственный» словарь, включающий как современную лексику с детально проработанной стандартизованной морфологией, так и все имеющиеся словари в качестве дополнительных специализированных приложений. Система такого масштаба требует концентрации соответствующих усилий для ее создания, развития и пополнения, в частности, и для перехода на следующих этапах к формированию многоязычной версии.

• «Тезаурус русского языка» по типу известных тезаурусов других языков. Значение подобного издания для нашей культуры, лингвистики и КТОЯ невозможно переоценить. Стоит упомянуть издательство Оксфордского университета, подготовившего к публикации «Исторический тезаурус Оксфордского словаря английского языка», самый большой подобный словарь в мире: 800 тысяч значений 600 тысяч слов, организованные в 354 категории и 230 тысяч подкатегорий. Идеографический словарь О. С. Баранова, представляющий собой практически единственный на сегодняшний день большой тезаурус русского языка, может послужить основой такого национального издания. Представляется, что при наличии средств за два — три года можно подготовить первый выпуск, причём сразу на трёх уровнях: полный, массовый и школьный.

4.3. *Речевые технологии:* Программа должна предусматривать исследования и разработки перспективных методов распознавания, синтеза, сжатия и идентификации русской речи.

Эти задачи включают разработку новой эффективной технологии полного транскрибирования речевого сигнала, в результате которой будут получены алгоритмы для систем обработки речи с характеристиками существенно лучшими и намного более точными, чем существующие в настоящее время. Эта часть Программы ориентируется прежде всего на естественную речь на русском языке и его диалектах по широкому спектру каналов речевой связи, в том числе введенную с микрофона, радиопередач, телефонных разговоров и т. п.

К основным областям, нуждающимся в разных формах речевых технологий относятся:

- Текстовые редакторы (ввод текста голосом),
- Сфера обслуживания (транспорт, торговля, call центры, др.),
- Управление техническими и бытовыми приборами, в частности, телевизорами и мобильными телефонами,
- Обработка речевой информации (извлечение данных, перевод, др.),
- Интерфейс «конечный пользователь — компьютер»,
- Документооборот (речевые материалы, тексты дискуссий, др.),
- Образование, прежде всего — обучение языкам,
- Каналы связи, в частности, мобильная связь, и многое другое.

5. Общая архитектура Программы

Очевидно, что целью Программы является не только интересующие разработчиков компоненты теории и технологии проблем обработки текста на естественном языке. Ее стратегическая цель — решение задач рынка, т. е. фронта соответствующих приложений.

Именно это определяет общую архитектуру Программы, которая, так или иначе, сводится к уровням

систем, покрывающим основные направления этого фронта. Эти уровни можно обозначить как:

- Базовые технологические составляющие КТОЯ, которые были кратко рассмотрены в п.4,
- Программные подсистемы, используемые как крупные строительные блоки систем обработки текста,
- Системы категории 1,
- Макросистемы,

В следующих разделах приводятся примеры компонентов этих уровней.

6. Программные компоненты систем КТОЯ

6.1. К уровню программных подсистем, используемых как крупные строительные блоки систем обработки текста, можно отнести компоненты КТОЯ, выделенные в следующие классы:

Корректоры текста. Технологии автоматического или автоматизированного исправления текста, включающие, в зависимости от выполняемых функций и их сложности, операции разного уровня, к которым относится:

- Коррекция орфографии,
- Коррекция пунктуации и синтаксиса,
- Коррекция стиля.

Автоматическая либо автоматизированная коррекция разного качества уже используется достаточно широко и может быть частью большинства приложений.

6.2. *Средства «предобработки».* Функции этой категории являются важными составляющими многих, если не большинства, приложений КТОЯ; они включают:

- Идентификацию языка входного текста,
- Нормализацию (унификацию) текста,
- Интеллектуальный OCR.

Идентификация языка сообщения может использоваться в речевых технологиях, возможно, наряду с предварительной классификацией речевого сообщения (определение жанра или ориентации ПО) как средство выбора соответствующего проблемного словаря

6.3. *Лингвистические процессоры* (анализаторы текста, парсеры). Программные модули, выполняющие основные функции КТОЯ, связанные, как правило, со следующими составляющими процесса обработки текста:

- Морфологический анализ,
- Синтаксический анализ,
- Анализ структуры текста.

В общем случае все эти компоненты лингвистических процессоров представляют собой тандемы <исполнитель, система правил>.

6.4. *Генерация (синтез) текста.* Перевод в текст формального представления информации. Таким представлением может быть:

- конструкция из типовых составляющих,
- синтаксическая структура текста (в основном, в МП),
- содержание, выраженное формальными средствами представления смысла (семантики и прагматики) генерируемого текста,
- сочетание этих уровней представления.

Приложениями, использующими генерацию текста в качестве подсистемы, являются:

- → машинный перевод,
- → диалоговые системы (генерация реплик компьютера),
- → подсистемы генерации сообщений Баз знаний, экспертных и диагностических систем, систем поддержки решений и др.,
- → подсистемы ответа на запрос для хранилищ данных (оформление ответа в виде ЕЯ-текста или комментарии к выдаваемым данным).

Кроме перечисленных приложений, имеется перспективный сектор КТОЯ, в котором основной функцией является именно генерация:

- → Типовых писем в деловой переписке,
- → Типовых документов в широком спектре от коротких готовых текстов, требующих подстановки нескольких параметров, до текстов большого размера (отчеты, типовые юридические документы, технические инструкции и др.)

Генерация текста должна использоваться во всех приложениях, связанных с использованием синтеза речи.

6.5. *Семантические процессоры* (извлечение содержания и понимание текста). В связи с повышением сложности и разнородности текстов растет необходимость в интеллектуализации технологий их обработки. К категории Семантических процессоров здесь отнесены КТОЯ, обеспечивающие в разной степени анализ и извлечение содержания обрабатываемого текста. В этом плане можно выделить четыре уровня интеллектуализации, связанные с учетом содержания документа, а именно:

- *Классификация текстов, формирование онтологий,*
- *Содержательная индексация текстов,*
- *Извлечение компонентов содержания,*
- *Понимание текста.*

Перечисленные типы процессов, так же как и лингвистические, представляют собой тандемы <исполнитель, система правил>.

- *Классификация текстов:* технология определения принадлежности текста к одной или нескольким категориям из заданного (возможно, расширяемого) набора категорий \ классов; как правило, строится на базе статистических методов. Формирование Онтологий является

наиболее продвинутой технологией создания содержательной структуры корпуса текстов, имеющей ключевое значение для остальных трех уровней интеллектуализации.

- *Содержательная индексация текстов:* технология обработки текста, основной функцией которой является повышение эффективности поиска в массиве текстов с помощью создания индекса, — структуры, сопоставляющей объектам поиска (текстовым компонентам) выделенные в них ключевые слова и словосочетания, маркирующие те объекты, в которых эти ключевые элементы присутствуют. Таким образом, ключевой элемент становится виртуальным адресом того подмножества текстовых компонентов, которые ему сопоставлены. В частности, индекс может служить базисом классификатора текстов.
- *Извлечение компонентов содержания* — один из основных элементов интеллектуализации технологий обработки текстов, роль которого быстро растет в связи с увеличением объемов и усложнением структуры информационного пространства, в частности, за счет роста сложности и разнородности самих документов. В зависимости от сложности задач, относимых к автоматическому или автоматизированному извлечению компонентов содержания из компьютеризированной информации (в основном, из текста, реже из БД и других форм данных) под извлечением содержания могут пониматься:
  - ⇨ Выделение значимых компонентов текста (даты, референты и т. п.),
  - ⇨ Выявление упоминаний о фактах и событиях из конкретного набора,
  - ⇨ Реферирование (выделение основного содержания),

Трудность создания эффективных технологий этого ряда возрастает с повышением сложности извлекаемого содержания. Более простые технологии используются практически во всех интеллектуальных системах обработки текста.

- *Понимание текста* — анализ содержания ЕЯ-текста и перевод его на формальный язык представления смысла в проекции на контекст анализируемого сообщения, т. е. *прагматику дискурса* и знания о *предметной области*, к которым оно имеет отношение. *Полное понимание* пока достигается только в экспериментальных системах и при анализе ЕЯ-запроса для *узких* (ограниченных) *предметных областей*.

Все четыре составляющие этой группы технологий являются необходимыми для развития речевых технологий, поскольку качество анализа речи непосредственно связано с выявлением ее содержания.

6.6. *Организация диалога «человек — компьютер».* Обмен репликами (команды, вопросы, сообщения) между пользователем и компьютером, направленный на решение конкретной проблемы: постановка задачи, уточнение ее условий, оценка результатов и т. п. С расширением сферы использования компьютеров и спектра решаемых задач диалог становится необходимой составляющей всей системы информатизации.

Сегодня возможности диалога ограничиваются в основном технологией меню, которая успешно обслуживает четко структурированные и ограниченные классы команд и запросов, но перестает быть эффективной при усложнении диалога и росте числа альтернатив.

Возможность *диалога на естественном языке* пока остается перспективой, близость которой для различных приложений определяется сложностью задачи и предметной области. ЕЯ-диалог требует уровня развития, на который соответствующие технологии выходят только в самое последнее время.

Освоение внеязыковых средств только начинается, но в будущем даст возможность значительно расширить возможности диалога, в частности, для учета эмоциональной составляющей пользователя (в частности, контроля состояния оператора в критических ситуациях), обучения иностранному языку, доступа к компьютеру слабослышащих или глухонемых людей, и т. п.

Области приложений диалога:

- Запрос к информационным системам,
- Интерфейсы к прикладным системам.

С развитием речевых технологий сфера устной речи сможет не только полностью покрыть сектор применения ЕЯ диалога (текст останется необходимым только там, где речь неприменима по техническим причинам), но и расширить его за счет тех приложений, где руки оператора заняты и\или набор текста непродуктивен .

7. ПРИЛОЖЕНИЯ

7.1. Приложения КТОЯ можно достаточно условно разделить по сложности на системы категории 1 и макропрограммы. К примерам систем первой категории отнесем следующие классы:

- Хранилища текстовых данных,
- Документооборот,
- Электронный архив,
- Электронная библиотека,
- Текстовый редактор.

Системы этой категории ориентированы на создание технологий, являющихся как самостоятельными продуктами, так и ключевыми компонентами макропрограмм.

7.2. Раздел макропрограмм объединяет наиболее крупные системы, ориентированные на создание стратегических продуктов и технологии КТОЯ:

- Корпоративные информационные системы,
- Интеллектуальные порталы,
- Системы машинного перевода,
- Интеллектуальные поисковые машины,
- Электронная торговля нового поколения и др.

Все макропрограммы, хотя и в разной степени, потенциально связаны с использованием речевых технологий.

8. Рынок технологий обработки текста и речи

В большинстве КТОЯ нуждаются практически все области деятельности, поскольку каждая из них требует таких систем как:

- Качественный документооборот,
- Тематические электронные архивы и информационные справочные системы, в частности, БД технической документации,
- Автоматическая содержательная обработка потоков текстовых сообщений,
- Аналитическая обработка текстовых данных в системах поддержки принятия решений,

Стоит привести несколько примеров областей широкого применения КТОЯ.

- Системы корпоративного управления любого уровня от госструктур и крупных корпораций до муниципалитетов и компаний среднего бизнеса нуждаются в массовом использовании КТОЯ, поскольку всем им требуется обработка текстовых документов, связанных со спецификой их деятельности.
- Системы образования (школы, ВУЗы, курсы и т. п.) включают те же компоненты, но специализированные в соответствии с особенностями данных секторов приложений. При этом для технологий этого направления важно развитие систем дистанционного обучения.
- Вся современная медицина базируется на специализированном документообороте, включающем:
  - ⇨ Ведение историй болезни,
  - ⇨ Выписку справок и рецептов,
  - ⇨ Те же электронные архивы и информационные справочные системы, связанные с медицинскими справочниками самого разного типа,
  - ⇨ Системы обучения, повышения квалификации и т. п.

Развитые медицинские учреждения нуждаются в экспертных и диагностических системах, разработка которых требует наличия баз знаний и аналитической обработки текстовых данных в системах поддержки принятия решений. Фактически поликлиника и больница должны превратиться в текстовый конвейер, автоматизированный во всех узлах контакта с человеком (пациент, врач, администратор).

- Системы двойного назначения для силовых министерств нуждаются в системах корпоративного управления, системах обучения и подготовки кадров, а также в таких технологиях как:
  - ⇨ Автоматическая содержательная обработка потоков текстовых сообщений,
  - ⇨ Эффективный интеллектуальный поиск,
  - ⇨ Машинный перевод,
  - ⇨ Комплекс речевых технологий,
  - ⇨ Совершенствование связи,
  - ⇨ Криминалистическая экспертиза и т. п.

Заключение

Оптимальной формой реализации Национальной Программы такого масштаба является формирование одного (или нескольких по направлениям) Консорциума на основе кооперации коммерческих фирм, научных коллективов, вузов и государственных структур, направленного на реализацию широкой национальной программы соответствующего профиля.

За последние десять лет автор участвовал в предварительной работе по подготовке нескольких аналогичных программ, которые, к сожалению, по разным причинам до реализации не доходили.

Соответствующие материалы, значительно более детальные по проработке, были использованы при подготовке этой статьи, в задачи которой входило дать представление о полноте и масштабе Программы, а не отразить ее с соответствующей полнотой и детализацией.

Хотелось бы надеяться, что время для реализации Программы пришло, поскольку без нее вряд ли можно планировать всерьёз такие национальные процессы как модернизация и инновация.