

3. Азия, Европа и США: черты сходства и различия

В последние годы двумя важнейшими движущими силами в развитии переводческой технологии были события 11 сентября и расширение ЕС.

3.1. Европейский Союз

Языковая ситуация и потребность в переводческих технологиях в значительной мере уже описаны в главе 2.3.

Резюмируем основные положения:

- Европейское общество является многоязычным, и это не случайно;
- Все официальные языки Европы обладают равными правами;
- Прямой и обратный перевод с этих 20 официальных языков, образующих 190 языковых пар, требует значительной работы;
- Любая мера, облегчающая перевод, содействовала бы общению между гражданами Европы и расширяла бы торговлю внутри Европейского Союза;
- Любой компании, работающей на рынке ЕС, приходится адаптировать свои товары к региональным рынкам, учитывая местные языки.

Европейская комиссия заинтересована в исследованиях в области переводческих технологий и в течение ряда лет финансирует несколько научных проектов по машинному переводу.

Оперативно получить информацию о проектах можно на сайте IST (Information Society Technologies)¹⁶. В неполном списке числятся LC-STAR¹⁷, MATCHPAD¹⁸, METIS и METIS-II¹⁹, NESPOLE!²⁰, TC-STAR_p и TC-STAR²¹, TQPRO²² и TransType2²³. Европейская комиссия играла решающую роль в создании необходимых лингвистических ресурсов, финансируя множество проектов в этой области исследований.

Согласно 6-й Рамочной программе, Европейский Союз тратит 135 миллионов евро на многомодальные интерфейсы и переводческие технологии, т.е. примерно по 15 миллионов евро в год на переводческие технологии.

Европейское научное сообщество по своей природе хорошо подходит для решения многоаспектной задачи создания переводческих технологий. В Европе немало крупных государственных научных учреждений мирового уровня, что позволяет делить работу между ними и двигаться разными путями. Принципы многоязычия и культурного многообразия в целом понимаются и принимаются как в научной, так и в деловой сферах и при этом обеспечиваются политической инфраструктурой. Уже 10 лет действует инфраструктура языковых

¹⁶ IST project search: under <http://www.cordis.lu/ist/projects/projects.htm>

¹⁷ LC-STAR: <http://www.lc-star.com/>

¹⁸ MATCHPAD: <http://www.systransoft.com/R&D/Matchpad/index.html>

¹⁹ METIS-II: <http://www.ilsp.gr/metis/>

²⁰ NESPOLE! : <http://nespole.itc.it>

²¹ TC-STAR: <http://www.tc-star.org/>

²² via the IST page: under <http://www.cordis.lu/ist/projects/projects.htm>

²³ TransType2: <http://tt2.atosorigin.es/>

ресурсов – *Европейская ассоциация языковых ресурсов* (European Language Resources Association – ELRA).

3.2. Соединенные Штаты Америки

3.2.1. Стратегическая роль переводческих технологий (ПТ)

В Соединенных Штатах стратегическая роль ПТ существенно иная, чем в Европе. США имеют громадный внутренний рынок с фактически одним языком. И это справедливо, несмотря на то что в стране значительное латиноамериканское население, говорящее на испанском языке. Однако отсутствует сильная экономическая потребность в обслуживании этого экономически слабого слоя населения, и нет законодательно оформленного требования о поддержке этого языка. По существу, Соединенные Штаты представляют собой рынок с единственным языком.

Мировое господство английского языка как иностранного во многих отношениях выгодно для США, но таит в себе страшный недостаток в плане национальной безопасности. При таких обстоятельствах перевод, особенно с других языков на английский, играет решающую роль в сборе информации со всего мира и обо всем мире, будь то общая информация или сведения, относящиеся к разведывательным операциям.

Интерес США к ПТ в значительной мере определяется факторами национальной безопасности.

3.2.2. Исследовательские программы

После нападения на Всемирный торговый центр 11 сентября 2001 года сразу стало ясно, что необходимость в

переводчиках и людях, владеющих различными языками, должна резко возрасти. Это была одна из причин, почему Соединенные Штаты запустили грандиозную исследовательскую программу по сбору информации из источников на разных языках. Ее бюджет – более 50 миллионов долларов в год.

Эти средства, в несколько раз превышающие суммы, расходуемые Европейской комиссией, а также сама программа будут оказывать стимулирующее действие на научное сообщество. Если некоторые научно-технические достижения будут соответствовать конкретным запросам Европы, то с другими дело обстоит иначе: сбор сведений для военной разведки по целому ряду аспектов отличен от задачи перевода для многоязычного общества. Более того, фактически одноязычная структура американского внутреннего рынка также вызывает законные сомнения в том, что именно американский рынок переводческих технологий создаст то, что нужно Европе.

Не вдаваясь в подробности этих исследовательских программ, можно в целом констатировать, что США обычно финансируют небольшое число крупных проектов, в то время как ЕС финансирует крупномасштабные программы, состоящие из более мелких проектов. Сегодня флагманским проектом является GALE (Global Autonomous Language Exploitation), который решает задачу извлечения информации из разноязычных текстовых и звуковых документов в *неограниченной* области (т.е. речь идет и о переводе устной речи). Осуществляется и ряд других переводческих проектов, например, TransTac, STR-DUST и ACTD.

НЕСЕКРЕТНО

Ведомость обоснования строк бюджета по НИОКР, испытаниям и оценке							Дата Февраль 2005 г.	
Статьи ассигнований/бюджета Все оборонные программы, прикладные исследования				Номенклатура статей Информационные и коммуникационные технологии				
ЗАТРАТЫ (млн. долл.)	2004 ф.г.	2005 ф.г.	2006 ф.г.	2007 ф.г.	2008 ф.г.	2009 ф.г.	2010 ф.г.	2011 ф.г.
Перевод с одного языка на другой	0,000	57,389	65,744	69,687	75,221	75,593	65,593	60,593

(U) Описание задачи:

(U) В проекте будет исследована и испытана мощная новая технология для обработки естественных языков, которая создаст принципиально новые возможности для удовлетворения самых различных нужд национальной безопасности. Благодаря этой технологии различные системы смогут: (а) автоматически использовать большие объемы устной речи и текстов на разных языках; (б) революционизировать взаимодействие человек – машина посредством английского и иностранных языков (как устных, так и письменных); (с) решать задачи расчета и принятия решений в напряженных, быстро меняющихся ситуациях; и (д) автономно подбирать, фильтровать, синтезировать и представлять релевантную информацию своевременно и в нужной форме. Данный элемент программы и проект сформированы в законопроекте по ассигнованиям Министерства обороны на 2005 финансовый год согласно намерению Конгресса. Финансирование на прошлый год выделено в бюджете РЕ 0602301Е, Проект ST-29, и указано как докладная записка в каждой из приведенных ниже программ.

(U) Выполнение программ / Запланированные программы:

Представление ситуации и презентация	2004 ф.г.	2005 ф.г.	2006 ф.г.	2007 ф.г.
	(10,870)	(11,500)	(11,616)	(14,387)

(U) Имеются две программы, предусматривающие непосредственный устный перевод с одного языка на другой:

Программа CAST (Compact Aids for Speech Translation) обеспечивает тактическому бойцу возможность перевода лицом к лицу в реальном времени в боевой обстановке и в ходе гуманитарных операций на чужой территории. Программа решает проблему быстрого действия и точности перевода в конкретных предметных областях. Первые опытные образцы программы CAST строились на простых словарях и словосочетаниях. Программа CAST позволяла прежде всего дать в распоряжение бойца систему быстрого одностороннего перевода (с английского на несколько иностранных языков). Главной из применяемых сегодня экспериментальных систем является Phraselator, разработанный Управлением перспективных исследовательских программ в области обороны. Система употреблялась в операции «Иракская свобода» и «Прочная свобода». Будущие версии сделают возможным более сложный, гибкий и гладкий перевод и перефразирование, они будут надежнее и более соответствовать нормальному диалогу между людьми.

Рис. 4. Бюджет Министерства обороны США на технологии перевода с языка на язык. Смета расходов Управления перспективных исследовательских программ в области обороны (DARPA) на 2005 финансовый год на исследования, разработки, испытания и оценку по всем оборонным программам. – Источник: DARPA.

Интервью с Джозефом Оливом, руководителем программы, DARPA

Д-р Джозеф Олив – руководитель программы Управления информационных технологий в DARPA. Сегодня в его портфеле одна крупная программа – GALE (Глобальная автономная эксплуатация языка).

Д-р Олив имеет более чем 30-летний опыт научно-исследовательской работы в центре Bell Laboratories. Он был лидером в синтезировании речи из текстовой информации и руководил командой мирового класса, занимавшейся системами компьютерного диалога и общением человека с компьютером. В роли Директора по изучению речи и Главного директора по технологиям в компании Lucent Speech Solutions он курировал окончательную доработку основных переводческих технологий Bell-Labs: «Распознавание речи», «Преобразование текста в речь» и «Опознавание говорящего».

Д-р Олив закончил Университет Чикаго с ученой степенью по физике (по специальности «Вычислительная атомная физика») и степенью магистра искусств по музыкальной композиции. По окончании университета он стал совмещать свой интерес к вычислениям и музыке, занявшись исследованиями в области акустики и обработки сигналов.

В 1974 г. д-р Олив получил от Национального фонда искусств грант на написание компьютерной оперы. В 1984 году он также получил от Bell Laboratories награду «Distinguished Member of Technical Staff».



Джозеф Олив,
руководитель программы,
Управление информационных
технологий (IPTO), DARPA,
Арлингтон, шт. Виргиния, США

Каковы научные направления и цели программы GALE?

Глобальная автономная эксплуатация языка (GALE) – это программа, которая будет выдавать нашим военным «отжатую», краткую и оперативную информацию. Поскольку источник может быть на иностранном языке, система GALE требует перевода и механизма дистилляции. Вводом может служить текст или устная речь, а выводом может быть полный перевод или «отжатый» ответ на запрос (не обязательно на естественном языке). Цель GALE – достичь такого уровня точности перевода и дистилляции, при котором система стала бы пригодной для военных целей. Программа GALE – продолжение двух программ DARPA по переводческим технологиям: EARS (программа транскрибирования) и TIDES (программа перевода, поиска, извлечения и синтеза информации). Предыдущие программы не были рассчитаны на построение сквозных систем перевода, но создали большой задел по переводческим технологиям.

Какова ваша позиция по вопросу о трансатлантическом научном сотрудничестве?

В Европе есть несколько групп, связанных с GALE. Они являются субподрядчиками при крупных командах, сформированных Главными исполнителями GALE, и сотрудничают внутри своих команд. В целом я считаю, что сотрудничество между американскими и европейскими/азиатскими учеными идет хорошо. Оно осуществляется на уровне правительств, корпораций и университетов.

Сколько государственных средств расходуется в США на переводческие технологии типа МП?

Общая программа DARPA по переводческим технологиям на 2006 год составляет примерно 50 миллионов долларов США. Я не думаю, что ее следует разбивать дальше, потому что переводческие технологии взаимосвязаны (или должны быть взаимосвязаны) и интегрированы, с тем чтобы добиться успеха в этой области. Я не располагаю информацией о расходах на ПТ других правительственных органов, но уверен, что DARPA не единственная организация, интересующаяся этой областью.

Насколько важны для США переводческие технологии, в особенности МП и родственные технологии?

Лично я считаю, что ПТ чрезвычайно важны для правительства США и военных. Нам необходимо общаться с союзниками и врагами и понимать их, а ведь многие из них не говорят по-английски. Также в связи с информационным взрывом нужно уметь находить важную информацию в «стоге сена».

Как организовано в США государственное финансирование исследований в области ПТ?

Мне не доводилось видеть в США никакого общего плана исследований по ПТ. Исследования поступают от различных правительственных органов и промышленности. Наблюдается широкое сотрудничество между государственными исследовательскими программами, корпоративными исследованиями и университетскими исследованиями, но оно носит скорее неформальный характер. Что касается DARPA, мы строим очень большие планы и ставим далеко идущие цели в отношении ПТ. И мы проводим периодические оценки для обеспечения прогресса. Мы работаем как путем сотрудничества, так и путем конкуренции. Наши команды нацелены на победу при оценках. Кроме того, мы увязали наши цели и оценки с полезностью технологии, а не просто с точностью измерения. ПТ имеют дело с языком, и в первую очередь их цель состоит в том, чтобы сохранить смысл в документе (устном или письменном).

Каков современный уровень развития переводческой технологии и с какими новыми научными проблемами придется столкнуться в этой области?

Технология МП в последние два-три года заметно усовершенствовалась. Этот прогресс в значительной мере обязан статистическим парадигмам в сочетании с методами оптимизации на основе BLEU. Впрочем, боюсь, что эта парадигма быстро исчерпывает свои возможности, если это уже не произошло. Я бы сказал, что машинный перевод находится сейчас примерно на полпути к тому, чем он должен быть. Хотелось бы увидеть появление новых идей для решения этой проблемы, пусть даже вначале результаты не будут конкурентоспособными. Необходимо, чтобы и МП, и технология автоматического распознавания речи взяли на вооружение комплексный подход, включив в себя другие методы ПТ, например информационный поиск, синтаксический анализ, извлечение и т.д. Хотя это тоже статистические методы, каждый из них работает по-своему, так что комбинированная система, использующая все эти технологии, могла бы улучшить результаты, что привело бы к решению задачи. Крайне важно еще теснее объединить автоматическое распознавание речи и машинный перевод, а не использовать первый лучший результат автоматического распознавания речи для запуска МП-компонента.

3.3. Восточная Азия

3.3.1. Английский как общепринятый язык в Восточной Азии

Учитывая разнообразие ее народов, культур и языков, Азия больше напоминает Европу, чем США. Но если Европа уже начала позиционировать себя как единое целое, то в Азии пока нет никакой ясной политики и нет ощущения себя как отдельной целостности. Страны общаются друг с другом в основном на английском языке, и широкое распространение в Азии английского как общепринятого языка говорит о том, что здесь существенно меньше спрос на прямой перевод языковых пар, например, тайский – японский. Преимущество заключается в том, что каждая страна в первую очередь должна гарантировать перевод на собственный язык с английского и с собственного языка на английский, когда на каждый язык приходится всего одна языковая пара, это упрощает ситуацию.

3.3.2. Азиатские языковые пары и растущее значение китайского языка

Однако использование в Азии английского как общепринятого языка общения имеет и обратную сторону. Возьмем в качестве примера задачу перевода между японским и китайским языками. Учитывая общие корни этих языков, слова в них имеют сходную этимологию. Было бы естественно воспользоваться этим обстоятельством, а также чертами культурного сходства. Применение английского

языка как посредника – не прямой способ, приводящий к ненужным осложнениям. Появилась новая тенденция: японские компании все чаще интересуются прямым переводом между японским и китайским языками. Это связано с бурным ростом китайской экономики и тем обстоятельством, что многие японские компании перебазировались в Китай. Интерес к переводу с китайского и на китайский растет и в мировом масштабе, и в Японии, и сегодня между Китаем и Японией более тесные отношения, чем в прошлом.

Английский является общепринятым языком в Азии. Растет значимость китайского языка.

Экономически сильная страна, язык которой имеет корни в древнекитайском, Корея поддерживает тесные связи и с Китаем и с Японией. Слова в японском и корейском языках в большинстве случаев имеют китайское происхождение, т.е. у них общая этимология, хотя в корейском языке не употребляются китайские иероглифы. Технические термины нередко вначале появляются в японском языке на основе китайских иероглифов, а затем воспринимаются в китайском языке. Это сходство в словарном запасе оказывается очень полезным при переводе; тем не менее эти три языка различаются произношением и совершенно различны в лингвистическом смысле. Учитывая эти общие культурные корни и тот факт, что торговля и контакты между этими тремя государствами сильнее,

чем с другими странами Азии, можно предположить, что на следующем этапе развития для трех языков – китайского, японского и корейского – будет непосредственно осуществляться прямой и обратный перевод, в то время как другие азиатские языки будут взаимодействовать главным образом через английский.

На следующем этапе для китайского, японского и корейского языков чаще будет осуществляться прямой и обратный перевод.

3.3.3. Исследовательские программы

В Японии пик исследований в области переводческих технологий пришелся на 80-е годы, когда на эти цели выделялись большие государственные средства, а компании проявляли большой интерес к проектам по машинному переводу, например к проекту EDR (*Electronic Dictionary Project*). Когда обнаружилось, что связанные с машинным переводом большие ожидания пока что не сбываются (вспомним европейский проект *EUROTRA*), интерес и государственное финансирование пошли на убыль. Сегодня имеются признаки возрождения интереса.

Учитывая господствующее положение английского языка, в исследованиях по переводу основной акцент делается на перевод с английского на родной язык и наоборот, что по существу приводит к тому, что научно-исследовательские программы осуществляются на национальном уровне. Межнациональные азиатские проекты было бы довольно трудно осуществить, поскольку в Азии для этого нет подходящей организации. Однако некоторые аспекты научной работы требуют международного сотрудничества, которое, как можно надеяться, будет осуществляться и поддерживаться азиатскими ассоциациями, например недавно основанной *Азиатской федерацией по обработке естественного языка* (*Asian Federation of Natural Language Processing*)²⁴.

Для стимулирования исследований в области перевода между языками Азии было бы весьма важно накопить языковые ресурсы для этих языковых пар. Учитывая требуемые затраты, а также необходимость совместного использования ресурсов, естественный путь к достижению этих целей – через международное сотрудничество. Следующим важным шагом на этом пути было бы создание специального органа оценки.

²⁴ <http://afnlp.org>

Интервью с Юн-ити Цудзии, директором Национального центра по разработке текстов в Манчестере и профессором в Манчестере и Токио

Профессор Юн-ити Цудзии в июле 2005 года был назначен директором Национального центра по разработке текстов (National Centre for Text Mining) и профессором по разработке текстов в Школе информатики при Манчестерском университете (Великобритания). Он также является профессором по обработке естественного языка на факультете компьютерных наук Токийского университета (Япония). В области обработки естественного языка он работает с 1976 года. Начав с машинного перевода, он расширил сферу своих исследований на грамматические формализмы для практических приложений обработки естественного языка, синтаксический анализ на базе HPSG-грамматики, извлечение информации и разумные ответы на вопросы. Его исследовательская группа по обработке естественного языка в Токийском университете недавно успешно применила глубокий синтаксический анализатор для получения семантических представлений всех рефератов в системе Medline (1,4 миллиарда слов).



Юн-ити Цудзии,
директор Национального центра
по разработке текстов,
Манчестер (Великобритания),
профессор Школы информатики
при Манчестерском
университете, профессор
факультета компьютерных наук
Токийского университета

Он считается признанным авторитетом в области разработки биотекстов, машинного перевода и обработки многоязычных текстов, и активно содействует паназиатскому сотрудничеству. Его приглашают давать консультации, читать лекции и делать основные доклады на многочисленных крупных конференциях как в области биоинформатики, так и в области обработки естественного языка. Он постоянный член ICCL (Международного комитета по вычислительной лингвистике) с 1992 года, был вице-президентом (2005 г.) и президентом (2006 г.) ACL (Ассоциации вычислительной лингвистики), президентом (2003–2005 гг.) IAMT (Международной ассоциации машинного перевода) и вице-президентом AFNLP (Азиатской федерации по обработке естественного языка).

В восьмидесятые годы машинный перевод был темой масштабных научно-исследовательских программ, но их результаты не оправдали ожиданий общественности. Как вы смотрите на эти работы сегодня?

Действительно, работы 80-х годов не оправдали ожиданий большого потенциального рынка. Как участник некоторых из этих проектов я был сказал, что мы испытали разочарование. Однако эти работы позволили создать прочную базу для исследований и разработок в области МП. Благодаря этим работам в Японии на рынке по-прежнему активно работают 6–7 поставщиков систем МП. Но важнее, как мне кажется, то, что наши тогдашние замыслы далеко опережали время. Иначе говоря, для достижения наших целей не было еще необходимых технологий. Полагаю, что теперь, когда мы располагаем технологиями, которых нам не хватало в то время, мы сумеем воскресить эту область. Например, одна и та же задача синтаксического анализа предложения с помощью сложных грамматических формализмов, на решение которой в то время требовались часы, сейчас занимает меньше секунды времени, когда мы выполняем ее с помощью нашей программы в Токийском университете.

Переводческие технологии, в том числе машинный перевод, должны представлять интерес для Азии. Так ли это воспринимается в разных странах?

К сожалению, нет. В отличие от Европы английский язык здесь давно считается единственным международным языком общения, и поэтому люди не думают, что имеет смысл работать с «местными языками» типа китайского, корейского, японского и т.д. Конечно, это неправильно, и люди начинают это понимать. Благодаря Интернету мы вдруг осознали, что существует громадный спрос на обработку «местных языков» и что наш местный язык на самом деле вовсе не местный.

В чем основные различия между азиатскими и европейскими языками? Насколько близки друг другу главные азиатские языки?

Все зависит от того, как мы определяем европейские и азиатские языки. Европейцы утверждают, что языки Европы разнообразны, и я соглашусь с этим. Но разнообразие азиатских языков колоссально и, как я думаю, намного превосходит в этом смысле европейские языки. Многие языки Индии, Среднего Востока, Малайского архипелага и др. совершенно не связаны с языками Дальнего Востока, китайским, корейским, японским. При этом китайский и японский/корейский языки, хотя и имеют общие словники благодаря долгой истории культурных обменов, принадлежат к совершенно разным языковым семьям.

Как бы вы охарактеризовали в целом ситуацию во всех странах Азии?

Я не подходящий человек для ответа на этот вопрос, но ситуация в азиатских странах опять же отличается разнообразием. Существует немало языков, для которых нет согласованных правил транскрипции, не говоря уже о стандартизации кодов символов. Впрочем, поскольку во многих странах, например Индии, Таиланде, Китае и т.д., наблюдается ускорение технологического прогресса, эти страны имеют более или менее сходный технологический уровень. Нам встречается немало интересных научных статей, опубликованных учеными из этих стран.

Наблюдается ли в Азии какая-либо международная деятельность в направлении координации научных исследований в этом секторе или даже сотрудничество в мировом масштабе?

Два года назад мы создали научную ассоциацию, Азиатскую федерацию по обработке естественного языка (AFNLP) с целью поощрять сотрудничество и координировать работы. Есть и другие инициативы. Но по сравнению с Европой мы еще очень далеки от идеала.

Какие следующие шаги нужно предпринять для содействия развитию ПТ и МП в Азии?

Необходимо более широкое участие государства. ЕС играет существенную роль в развитии НИР в области ПТ и МП. До конца 80-х годов Япония играла ведущую роль не только в технологии, но и в финансировании регионального сотрудничества в этой сфере. С тех пор ситуация резко изменилась. Мы видим несколько стран, которые могут внести не только интеллектуальный, но и финансовый вклад в развитие отрасли. Однако у нас нет никакого паназиатского правительственного органа, который координировал бы эту деятельность. Научное сообщество готово к сотрудничеству, но нам нужны, например, финансирующие органы, которые поддерживали бы все азиатские проекты.

3.4. Индия

Можете ли вы представить себе страну с таким же языковым разнообразием, как в объединенной Европе? Так вот, в Индии²⁵ с ее 35 штатами насчитывается 22 языка, которые считаются официальными, то есть утверждены конституцией. Мало того, каждый язык имеет в среднем по 20 диалектов. Кроме этого, все языки принадлежат к трем разным языковым семьям, и к тому же имеются разные письменности. Государственный язык – хинди, но в индийской конституции записано, что английский язык тоже может использоваться для официальных целей. В городах люди часто знают три языка: язык штата (например, хинди, бенгали и т.д.), хинди (официальный язык Индии) и английский. Примерно 30% населения говорит на хинди, и примерно 5% достаточно хорошо владеет английским языком.

Все официальные документы в столицах штатов должны быть на трех языках (на английском, хинди и языке штата). Самое распространенное направление перевода – с английского на хинди и с английского или хинди на соответствующие языки штатов. Но, как и в других странах, большинство переводов выполняется людьми, а их недостаточно, чтобы

удовлетворить спрос. Если говорить о технических средствах, то система «переводческая память» не имеет широкого применения, но перевод часто поддерживается электронными словарями и такими средствами, как морфологические анализаторы.

И общество, и правительство проявляют интерес к автоматическому переводу, и государство финансирует индийские исследования по машинному переводу. Международные компании типа IBM, Microsoft, Google и Yahoo инвестируют в МП, хотя он пока еще не работает с индийскими языками. Поскольку эти компании, как правило, используют статистический подход, существует множество параллельных фондов²⁶, но они не являются общедоступными. Начиная с 2006 года этот недостаток будет устранен, когда возникнет новый государственный орган – LDC-IL²⁷ (Консорциум лингвистических данных по индийским языкам – *Linguistic Data Consortium for Indian Languages*).

3.5. Экономические граничные условия

Экономические граничные условия применительно к переводческим технологиям различны в Европе и в США. На самом деле рыночные условия для систем распознавания

²⁵ Выражаю благодарность профессору д-ру Пушпаку Бхаттачария (Индийский технологический институт, Мумбай) за предоставленную информацию о ситуации в Индии.

²⁶ Еще один параллельный фонд – собрание текстов на двух версиях языка вместе с информацией о том, какие предложения связаны друг с другом. (Например, первые два предложения в языке А могут соответствовать первым трем предложениям в языке В и т.д.).

²⁷ <http://www.ciiilcorp.net/ldcil.htm>

речи и машинного перевода в США и ЕС прямо противоположны.

Вначале давайте рассмотрим автоматическое распознавание речи. В данный момент самая интересная с экономической точки зрения сфера использования автоматического распознавания речи и диалоговой технологии – самообслуживание клиентов, иначе говоря, автоматизация центров обработки вызовов, обслуживаемых людьми. Что касается этого рынка и этой сферы, Европа отличается от США в двух отношениях. Во-первых, если в США широко применяются так называемые интерактивные автоответчики (IVR)²⁸, то в Европе они не так распространены. Американцы уже довольно давно пользуются этими системами, нередко несколько утомительными, и в целом были вполне довольны переходом с DTMF-сервиса (двухтонального многочастотного набора номера) и жестких меню на речевые диалоги²⁹. В Европе, где в некоторых странах информационно-справочные службы бесплатны и предоставляют весьма высокий уровень обслуживания, введение автоматических систем часто рассматривается звонящими как шаг назад. К тому же США – это большой и достаточно однородный рынок с одним языком, гораздо больший, чем любой одноязычный ев-

ропейский рынок. В результате в США сильнее эффект масштаба, поскольку разработка устройства речевого диалога требует значительных усилий. С учетом этого неудивительно, что автоматическое распознавание речи больше прижилось в США, чем в ЕС.

Прямо противоположны экономические условия для машинного перевода: они трудны для США, но очень благоприятны для Европы в отношении и устной речи, и машинного перевода. Это связано с тем, что в Европе говорят на разных языках, в то время как любая система, используемая для громадного внутреннего американского рынка, охватывала бы лишь малую долю всей деятельности. Заниматься коммерцией в Европе значит быть многоязычным.

Благоприятные экономические условия для европейских компаний, предоставляющих услуги МП.

Также имеются мощные экономические силы, поддерживающие переводческую отрасль в Азии, но вследствие господствующего положения английского языка местный рынок, а также перспективы для переводческой индустрии меньше, чем в Европе. Учитывая текущую ситуацию и положение компаний, мы бы предположили, что при

²⁸ IVR означает «интерактивный речевой ответ». В классических системах IVR после нажатия звонящим кнопок проигрываются речевые сообщения.

²⁹ В часто цитируемом отчете «Nuance Communications - Market Research: Nuance Speech User Scorecard» (май 2000 г.) говорится, что у 87% респондентов высокая общая удовлетворенность распознаванием речи, намного выше, чем удовлетворенность DTMF-системами. Хотя по крайней мере этот результат обоснован и совпадает с собственным опытом автора, следует проявлять осторожность, ибо первоначальный отчет больше не присутствует ни на сайте компании Nuance, ни в той части Интернета, которая доступна для Google.

обычных обстоятельствах Европа будет играть здесь ведущую роль.

3.6. Место действия для Европы

Обобщая сказанное, стратегическая роль ПТ, а также ситуация в трех больших регионах выглядит следующим образом.

Европа: Языковая технология настолько необходима с экономической, политической и культурной точек зрения. Преодоление языкового барьера будет способствовать общению и развитию экономики. Хотя ПТ уже являются предметом серьезных исследований в Европе, учитывая стратегическое значение этой технологии для Европы, она должна получить гораздо более высокий приоритет на повестке дня НИР.

США: Применение ПТ диктуется преимущественно военными сообра-

жениями и борьбой с терроризмом. На исследования и технологию ПТ выделяются сегодня очень значительные средства.

Азия: Главный приоритет – перевод с английского и на английский. Отсутствие общей политической идентификации и инфраструктуры мешает Восточной Азии взять на себя ведущую роль. Большой спрос на перевод наблюдается в Индии.

Что касается разработок и коммерческого использования, Европа ощущает в этом острую потребность и в то же время находится в выгодном положении. Учитывая общий характер данной технологии, ее промышленное внедрение возможно и в других регионах мира. С другой стороны, нельзя рассчитывать на то, что какие-то внешние поставщики удовлетворят наши потребности.

Интервью с Жозефом Марьяни, директором Департамента Министерства научных исследований Франции

Научная деятельность Жозефа Марьяни связана с переводческими технологиями, многомодальным общением человек – машина, распознаванием речи, языковыми ресурсами и вопросами оценки.

Он был президентом Европейской ассоциации языковых ресурсов (ELRA), президентом Европейской (ныне Международной) ассоциации речевой коммуникации (ISCA), членом правления Европейской сети по языку и речи (ELSNET), и координатором франкоязычной сети FRANCIL.

Д-р Марьяни был директором LIMSI и руководителем его Отдела коммуникации человек – машина (1989–2001 гг.), членом Научного совета Национального центра научных исследований (CNRS), председателем Консультативного совета по информационной науке и технологии CNRS и членом Комитета по оценке INRIA.

С 2001 года является директором Отдела информационно-коммуникационных технологий в Министерстве научных исследований Франции, где курирует исследовательские программы в области телекоммуникаций, программотехники, мультимедиа и нанотехнологий, в том числе конкретную программу по переводческим технологиям.



Жозеф Марьяни,
директор Департамента
информационно-
коммуникационных технологий
Министерства научных
исследований Франции
и старший научный сотрудник
LIMSI-CNRS, Париж (Франция)

По поводу Европы, США и Азии: в чем вы видите сходство между ними и в чем различие?

Соединенные Штаты с их преимущественно одноязычным внутренним рынком рассматривают многоязычие в первую очередь с точки зрения обороны и безопасности. Там все понимают по-английски, но с иностранными языками у них туго! В Азии, как и в Европе, множество языков, и там тоже ощущается коммерческая потребность в многоязычии, но это все же не общий рынок, как в ЕС. Для Европы, где в 25 странах – членах ЕС говорят более чем на 20 языках, многоязычие крайне важно. Помимо экономического аспекта, который делает многоязычие совершенно необходимым, существуют также политический, культурный и общественный аспекты. Это делает нашу ситуацию совершенно особенной.

Как вы относитесь к сотрудничеству с США или Азией?

Международное сотрудничество в совокупности со здоровой конкуренцией полезно для научно-технического прогресса, и я ценю и поддерживаю те инициативы, где наши страны объединяют усилия в науке и инфраструктуре для решения сложных проблем в области ПТ. С другой стороны, международное сотрудничество не может заменить программу удовлетворения наших собственных потребностей. Что касается многоязычия в Европе, то лидером должен быть ЕС – никто другой не может и не будет этого делать за нас.

Что нужно для того, чтобы укрепить позиции Европы?

Несмотря на довольно значительный объем работы, сделанной в прошлом, полагаю, что уровень финансирования научных исследований и технологий пока не соответствует масштабу вызова. Хотя Комиссар по вопросам образования, культуры и многоязычия Иен Фигель недавно подчеркнул значение многоязычия для Европы, переводческие технологии фигурируют в виде крошечной доли работ в области ИКТ в проекте 7-й Рамочной программы, потерявшись в конце технологического направления «Моделирование, визуализация, взаимодействие и смешанная реальность». А темы, отобранные для возможных крупных мероприятий в рамках Статьи 169, это «Научные исследования в Балтийском море», «Помощь престарелым» или «Метрология», каждая из которых весьма достойна, но, по моему мнению, имеет меньшее стратегическое значение для Европы, чем проблемы языка. Переводческие технологии – это ведь такая тема, которая лучше всего согласуется с принципом совместных действий в европейском масштабе, потому что идея охватить все технологии и все европейские языки слишком грандиозна, чтобы Европейская комиссия могла осуществить ее в одиночку. Ее вполне могли бы совместно разрабатывать страны – члены ЕС со своими языками и собственными программами. Если государства – члены должны в первую очередь учитывать то, что характерно для их языка (или языков), например языковые ресурсы (речевые и текстовые массивы, словари) и адаптированные к конкретному языку технологии, то Европейская комиссия могла бы заниматься в первую очередь аспектами, независимыми от конкретного языка, общей координацией работ, разработкой и оценкой базовых технологий и стандартами.

Европа все еще нуждается в инфраструктуре для оценивания переводческих технологий, сравнимой с Национальным институтом стандартов и технологий (NIST) в США, и финансирующие органы при отборе проектов должны учитывать результаты оценки опорных технологий, с тем чтобы не допускать финансирования разработки продуктов, явно требующих более совершенной технологии.

Это было бы дополнением к Европейской ассоциации языковых ресурсов (ELRA), только что отметившей свое десятилетие.

Да, действительно. Ситуация с языковыми ресурсами и оценкой переводческих технологий стала намного лучше благодаря таким постоянно действующим органам, как NIST и LDC (Консорциум по лингвистическим данным) в США или ELRA в Европе. Но проблемы, стоящие перед Европой, требуют большего. Я выступаю за создание Агентства переводческих технологий на уровне Европейской комиссии. Это должна быть постоянная структура, которая координировала бы работы Европейской комиссии и государств – членов ЕС по созданию необходимых переводческих технологий для многоязычной Европы и которая вывела бы Европу на передний край ПТ во всем мире. Тем самым она использовала бы языковые проблемы Европы в своих интересах.