Машинные переводчики

- **Маши́нный перево́д** — процесс перевода текстов (письменных, а в идеале и устных) с одного естественного языка на другой с помощью специальной компьютерной программы. Так же называется направление научных исследований, связанных с построением подобных систем.
- Следующее разделение машинных переводов основано на лекциях *Лари Чайлдса*, проведенных в рамках Международной Конференции по Техническим Коммуникациям 1990 года:

  - полностью автоматический перевод;
  - автоматизированный машинный перевод при участии человека;
  - перевод, осуществляемый человеком с использованием компьютера.

- **Полностью автоматизированный машинный перевод**
- Этот вид машинного перевода и подразумевается большинством людей, когда они говорят о машинном переводе. Смысл здесь прост: в компьютер вводится текст на одном языке, этот текст обрабатывается и компьютер выводит этот же текст на другом языке. К сожалению, реализация такого вида автоматического перевода сталкивается с определенными препятствиями, которые еще предстоит преодолеть.

- **Основной проблемой** является сложность языка как такового. Возьмем, к примеру, значения слова "can". Помимо основного значения модального вспомогательного глагола, у слова "can" имеется несколько официальных и жаргонных значений в качестве существительного: "банка", "отхожее место", "тюрьма". Кроме этого, существует архаичное значение этого слова - "знать или понимать". Если предположить, что у выходного языка для каждого из этих значений имеется отдельное слово, каким образом может компьютер их различить?
- Как оказалось, определенные успехи были достигнуты в сфере разработки программ перевода, различающих смысл основываясь на контексте. Более поздние исследования при анализе текстов опираются больше на теории вероятности. Тем не менее, полностью автоматизированный машинный перевод текстов с обширной тематикой все еще является невыполнимой задачей.

- **Автоматизированный машинный перевод при участии человека.**
- Этот вид машинного перевода теперь вполне осуществим. Говоря о машинном переводе при участии человека, обычно подразумевают редактирование текстов как до, так и после их обработки компьютером. Люди-переводчики изменяют тексты так, чтобы они были понятны машинам. После того, как компьютер сделал перевод, люди опять-таки редактируют грубый машинный перевод, делая текст на выходном языке правильным. Помимо такого порядка работы, существуют системы МП, во время перевода требующие постоянного присутствия человека-переводчика, помогающего компьютеру делать перевод особенно сложных или неоднозначных конструкций.

- Машинный перевод с помощью человека применим в большей степени к текстам с ограниченным вокабуляром узко-ограниченной тематики.
- Экономичность использования машинного перевода с помощью человека - вопрос все еще спорный. Сами программы обычно достаточно дорогостоящи, а для работы некоторых из них требуется специальное оборудование. Предварительному и последующему редактированию необходимо обучаться, да и работа эта не из приятных. Создание и поддержание в рабочем состоянии баз данных слов - процесс трудоемкий и зачастую требует специальных навыков. Однако для организации, переводящей большие объемы текстов в четко-определенной тематической сфере, машинный перевод с помощью человека может оказаться достаточно экономичной альтернативой традиционному человеческому переводу.

- **Перевод, осуществляемый человеком с использованием компьютера**
- При этом подходе человек-переводчик ставится в центр процесса перевода, в то время как программа компьютера расценивается в качестве инструмента, делающего процесс перевода более эффективным, а перевод - точным. Это обычные электронные словари, которые обеспечивают перевод требуемого слова, возлагая на человека ответственность за выбор нужного варианта и смысл переведенного текста. Такие словари значительно облегчают процесс перевода, но требуют от пользователя определенного знания языка и затрат времени на его осуществление. И все же сам процесс перевода значительно ускоряется и облегчается.

- Среди систем, помогающих переводчику в работе, важнейшее место занимают так называемые системы **Translation Memory (TM)**. Системы TM представляют собой интерактивный инструмент для накопления в базе данных пар эквивалентных сегментов текста на языке оригинала и перевода с возможностью их последующего поиска и редактирования. Эти программные продукты не имеют целью применение высокоинтеллектуальных информационных технологий, а наоборот, основаны на использовании творческого потенциала переводчика. Переводчик в процессе работы сам формирует базу данных (или же получает ее от других переводчиков или от заказчика), и чем больше единиц она содержит, тем больше отдача от ее использования.
- Вот список наиболее известных систем TM:
- - Transit швейцарской фирмы Star,
- - Trados (США),
- - Translation Manager от IBM,
- - Eurolang Optimizer французской фирмы LANT,
- - DejaVu от ATRIL (США),
- - WordFisher (Венгрия).

- Системы ТМ позволяют исключить повторный перевод идентичных фрагментов текста. Перевод сегмента осуществляется переводчиком только один раз, а затем каждый следующий сегмент проверяется на совпадение (полное или нечеткое) с базой данных, и, если найден идентичный или похожий сегмент, то он предлагается в качестве варианта перевода.
- В настоящее время ведутся разработки по усовершенствованию систем ТМ. Например, ядро системы Transit фирмы Star реализовано на основе технологии нейронных сетей.

- Несмотря на широкий ассортимент систем ТМ, они имеют несколько общих функций:
- - Функция сопоставления (Alignment). Одно из преимуществ систем ТМ – это возможность использования уже переведенных материалов по данной тематике. База данных ТМ может быть получена путем посегментного сопоставления файлов оригинала и перевода.
- - Наличие фильтров импорта – экспорта. Это свойство обеспечивает совместимость систем ТМ с множеством текстовых процессоров и издательских систем и дает переводчику относительную независимость от заказчика.

- - Механизм поиска нечетких или полных совпадений. Именно этот механизм и представляет собой основное достоинство систем ТМ. Если при переводе текста система встречает сегмент, идентичный или близкий к переведенному ранее, то уже переведенный сегмент предлагается переводчику как вариант перевода текущего сегмента, который может быть подкорректирован. Степень нечеткого совпадения задается пользователем.
- - Поддержка тематических словарей. Эта функция помогает переводчику придерживаться глоссария. Как правило, если в переводимом сегменте встречается слово или словосочетание из тематического словаря, то оно выделяется цветом и предлагается его перевод, который можно вставить в переводимый текст автоматически.

- - Средства поиска фрагментов текста. Этот инструмент очень удобен при редактировании перевода. Если в процессе работы был найден более удачный вариант перевода какого-либо фрагмента текста, то этот фрагмент может быть найден во всех сегментах ТМ, после чего в сегменты ТМ последовательно вносятся необходимые изменения.
- Конечно, как и любой программный продукт, системы ТМ имеют свои достоинства и недостатки, и свою область применения. Однако в отношении систем ТМ, основным недостатком является их дороговизна.
- Особенно удобно использовать системы ТМ при переводе таких документов, как руководства пользователя, инструкции по эксплуатации, конструкторская и деловая документация, каталоги продукции и другой однотипной документации с большим количеством совпадений.

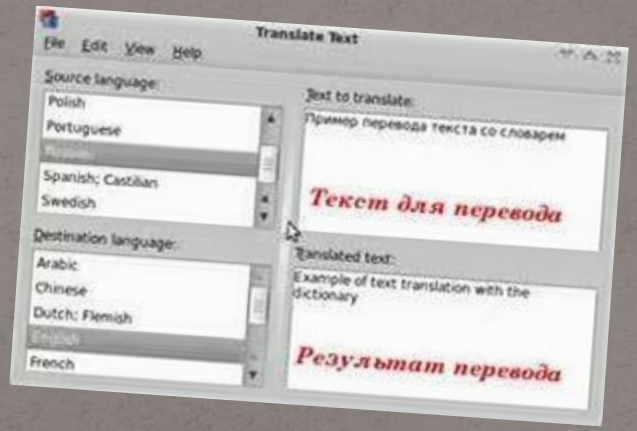
- При машинном переводе наиболее распространенной является следующая **последовательность формальных операций**, обеспечивающих анализ и синтез:
- 1. На первом этапе осуществляется ввод текста и поиск входных словоформ (слов в конкретной грамматической форме, например дательного падежа множественного числа) во входном словаре (словаре языка, с которого производится перевод) с сопутствующим морфологическим анализом, в ходе которого устанавливается принадлежность данной словоформы к определенной лексеме (слову как единице словаря). В процессе анализа из формы слова могут быть получены также сведения, относящиеся к другим уровням организации языковой системы.

- . Следующий этап включает в себя перевод идиоматических словосочетаний, фразеологических единств или штампов данной предметной области (например, при англо-русском переводе обороты типа in case of, in accordance with получают единый цифровой эквивалент и исключаются из дальнейшего грамматического анализа); определение основных грамматических (морфологических, синтаксических, семантических и лексических) характеристик элементов входного текста (например, числа существительных, времени глагола, синтаксических функций словоформ в данном тексте и пр.), производимое в рамках входного языка; разрешение омографии (конверсионной омонимии словоформ – скажем, англ. round может быть существительным, прилагательным, наречием, глаголом или же предлогом); лексический анализ и перевод лексем. Обычно на этом этапе однозначные слова отделяются от многозначных (имеющих более одного переводного эквивалента в выходном языке), после чего однозначные слова переводятся по спискам эквивалентов, а для перевода многозначных слов используются так называемые контекстологические словари, словарные статьи которых представляют собой алгоритмы запроса к контексту на наличие/отсутствие контекстных определителей значения.

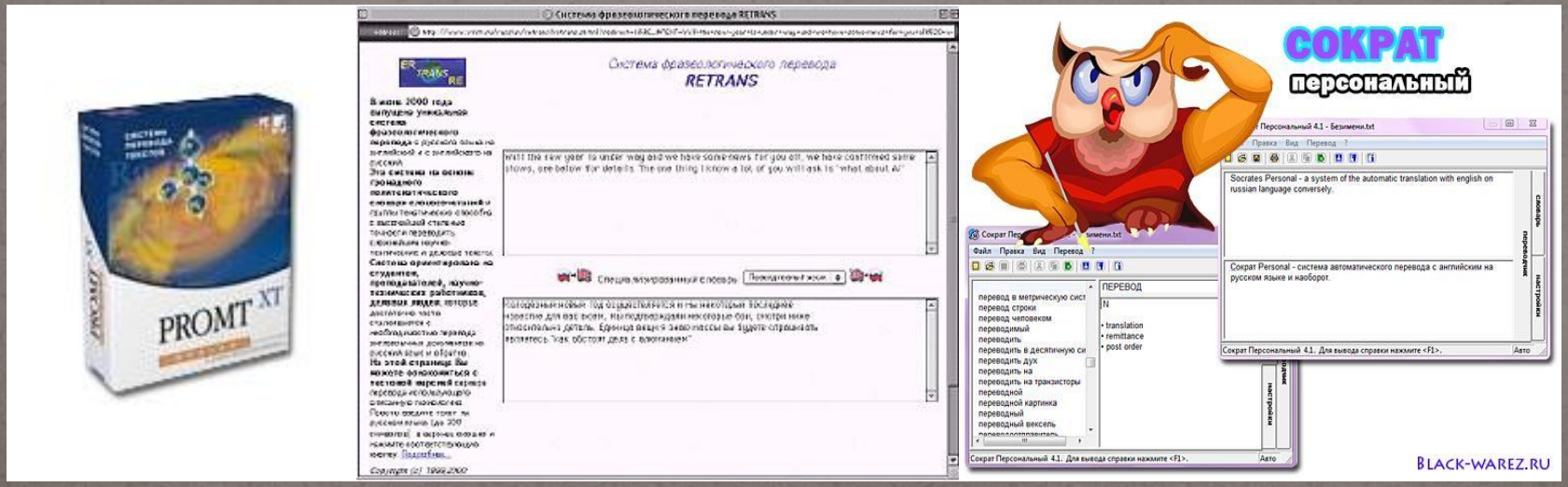
- 3. Окончательный грамматический анализ, в ходе которого доопределяется необходимая грамматическая информация с учетом данных выходного языка (например, при русских существительных типа сани, ножницы глагол должен стоять в форме множественного числа, несмотря на то, что в оригинале может быть и единственное число).
- 4. Синтез выходных словоформ и предложения в целом на выходном языке.
- В зависимости от особенностей морфологии, синтаксиса и семантики конкретной языковой пары, а также направления перевода общий алгоритм перевода может включать и другие этапы, а также модификации названных этапов или порядка их следования, но вариации такого рода в современных системах, как правило, незначительны. Анализ и синтез могут производиться как пофразно, так и для всего текста, введенного в память компьютера; в последнем случае алгоритм перевода предусматривает определение так называемых анафорических связей (такова, например, связь местоимения с замещаемым им существительным – скажем, местоимения им со словом местоимения в самом этом пояснении в скобках).

- **В настоящее время существует две концепции развития систем МП**:

1. Модель «большого словаря со сложной структурой», которая заложена в большинство современных программ-переводчиков;
2. Модель «смысл-текст», впервые сформулированная А.А. Ляпуновым, но пока что не реализована ни в одном коммерческом продукте.



- На сегодняшний день наиболее известны такие системы машинного перевода, как
- - PROMT 2000/XT компании PROMT;
- - Retrans Vista компаний Vista и Advantis;
- - Сократ – набор программ компании Арсеналъ.



- В большинстве случаев при работе над проектом применение систем МП не оправдано, поскольку:
- - Системы МП не дают приемлемого качества выходного текста. Более высокого качества можно добиться с помощью предварительной настройки системы (продукты серии PROMT XT предоставляют пользователю множество возможностей для этого), что совершенно неприемлемо при небольших объемах переводимого текста, и/или путем последующего редактирования, а это только замедляет работу, если переводчик использует слепой метод печати.
- - Системы МП не гарантируют соблюдения единства терминологии, особенно при работе коллектива переводчиков над большим проектом. Вернее, могут гарантировать при условии очень внимательного обращения с пользовательскими словарями, а на это не всегда стоит рассчитывать.

Система МП Retrans Vista

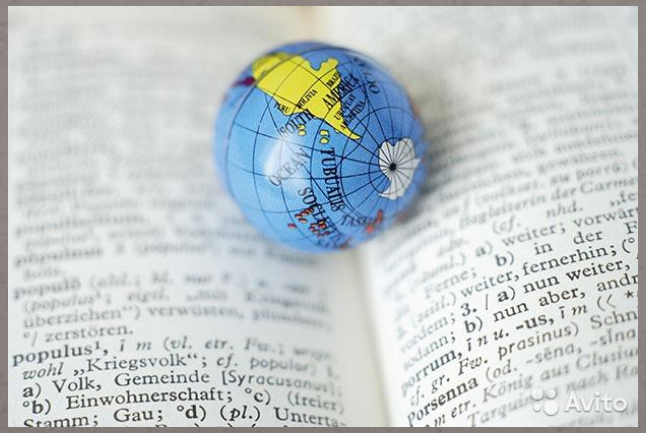
- **Важнейшими принципами являются следующие**:
- 1. Основными единицами языка и речи, которые, прежде всего, следует включать в машинный словарь, должны быть фразеологические единицы (словосочетания, фразы). Отдельные слова также могут включаться в словарь, но они должны использоваться только в тех случаях, когда не удается осуществить перевод, опираясь только на фразеологические единицы.
- 2. Наряду с фразеологическими единицами, состоящими из непрерывных последовательностей слов, в системах машинного перевода следует использовать и так называемые "речевые модели" - фразеологические единицы с "пустыми местами", которые могут заполняться различными словами и словосочетаниями, порождая осмысленные отрезки речи.

- 3. Реальные тексты, независимо от их принадлежности к той или иной тематической области, обычно бывают политематическими, если они имеют достаточно большой объем. Поэтому машинный словарь, предназначенный для перевода текстов даже только из одной тематической области, должен быть политематическим, а для перевода текстов из различных предметных областей - тем более. Он должен создаваться, прежде всего, на основе автоматизированной обработки двуязычных текстов, являющихся переводами друг друга, и в процессе функционирования систем перевода.
- 4. Наряду с основным политематическим словарем большого объема, в системах фразеологического машинного перевода целесообразно использовать также набор небольших по объему дополнительных тематических словарей. Дополнительные словари должны содержать только ту информацию, которая отсутствует в основном словаре (например, информацию о приоритетных переводных эквивалентах словосочетаний и слов для различных предметных областей).

- В ВИНИТИ РАН были построены две системы фразеологического машинного перевода:
- 1) система русско-английского перевода (RETRANS)
- 2) система англо-русского перевода (ERTRANS).

- Обе системы имеют одинаковую структуру и примерно одинаковые объемы машинных словарей. Поэтому мы рассмотрим только первую систему.
- **Система RETRANS имеет следующие характеристики**:
- 1. Область применения, назначение, функциональные возможности. Система предназначена для автоматизированного перевода научно-технических текстов с русского языка на английский. Русско-английский политематический машинный словарь системы содержит терминологию по естественным и техническим наукам, экономике, бизнесу, политике, законодательству и военному делу. В частности, он содержит термины и фразеологические единицы по следующим тематическим областям: Машиностроение, Электротехника, Энергетика, Транспорт, Аэронавтика. Космонавтика, Робототехника, Автоматика и Радиоэлектроника, Вычислительная Техника, Связь, Математика, Физика, Химия, Биология, Медицина, Экология, Сельское Хозяйство, Строительство и Архитектура, Астрономия, География, Геология, Геофизика, Горное Дело, Металлургия и др.

- Перевод текстов может осуществляться в автоматическом и в диалоговом режимах.
- 2. Объем политематического машинного словаря: более 1.300.000 словарных статей; 77 процентов из них составляют словосочетания длиной от двух до семнадцати слов. Объем дополнительных машинных словарей (для настройки системы на различные тематические области) - более 200.000 словарных статей.



Система МП PROMT XT

- В основу программных продуктов компании PROMT поставлено **решение следующих фундаментальных проблем**:

Во-первых, всем ясно, что чем больше словарь, тем лучше перевод, значит, первая проблема - проблема создания больших словарей для систем.

Во-вторых, ясно, что система должна переводить такие предложения: ПРИВЕТ, КАК ДЕЛА? Значит, еще одна проблема - научить систему распознавать устойчивые обороты.

В-третьих, понятно, что предложение для перевода пишется по определенным правилам, по определенным правилам переводится, а значит есть еще одна проблема: записать все эти правила в виде программы.

- В системах семейства PROMT разработано практически уникальное по полноте морфологическое описание для всех языков, с которыми системы умеют обращаться. Оно содержит 800 типов словоизменений для русского языка, более 300 типов, как для немецкого, так и для французского языка, и даже для английского, который не принадлежит к флективным языкам, выделено более 250 типов словоизменений. Множество окончаний для каждого языка хранится в виде древесных структур, что обеспечивает не только эффективный способ хранения, но и эффективный алгоритм морфологического анализа.
- Кроме того, используемая модель морфологии позволила разработать экспертную систему для пользователя - создателя словаря. Эта система фактически автоматизирует процедуру выделения основы и определения типа словоизменения при вводе новых словарных статей.

- Во многих системах МП в прошлом (как, впрочем, и сейчас) словарное описание и описание алгоритмов рассматривались как стороны одной проблемы, но решение, как правило, искалось в ограничении рассматриваемого мира, либо грамматического, либо семантического. Например, на основе признака "принадлежность к части речи" описывалась грамматика такого типа:
- именная группа - это существительное
- именная группа - это прилагательное + именная группа
- глагольная группа - это глагол + именная группа
- предложение - это именная группа + глагольная группа

- Именно из таких проектов появились системы перевода, которые сейчас предлагаются конечному пользователю. Это и Power Translator (компания Globalink) и Language Assistant (компания MicroTac) и TRANSEND (компания Intergraph).
- Системы семейств STYLUS и PROMT - не исключение, поскольку многие специалисты компании PROMT имели опыт работы в такого типа проектах. Однако при разработке систем PROMT впервые был применен фактически революционный подход, который и позволил получить впечатляющие результаты. Системы перевода семейства PROMT - это системы, спроектированные на основе не лингвистических, а кибернетических методов.

- Вместо принятого лингвистического подхода, предполагающего выделение последовательных процессов анализа и синтеза предложения, в основу архитектуры систем было положено представление процесса перевода как процесса с "объектно-ориентированной" организацией, основанной на иерархии обрабатываемых компонентов предложения. Это позволило сделать системы PROMT устойчивыми и открытыми.
- Кроме того, такой подход дал возможность применения различных формализмов для описания перевода разных уровней. В системах работают и сетевые грамматики, близкие по типу к расширенным сетям переходов, и процедурные алгоритмы заполнения и трансформаций фреймовых структур для анализа сложных предикатов.

Спасибо за внимание!