Системы автоматического (машинного) перевода текста

История, основные сведения, описание.

Лекция №13

Лингвистическое обеспечение АСОИУ

К.т.н., доцент Филиппович Анна Юрьевна

Автоматический (машинный) перевод

- ***Машинный перевод*** (МП) – автоматический перевод, перевод текстов с одного языка на другой с помощью автоматических устройств.
- ***Машинный перевод*** – выполняемое на компьютере действие по преобразованию текста на одном естественном языке в эквивалентный по содержанию текст на другом языке, а также результат такого действия.

«Лингвистический арифмометр» Смирнова-Троянского

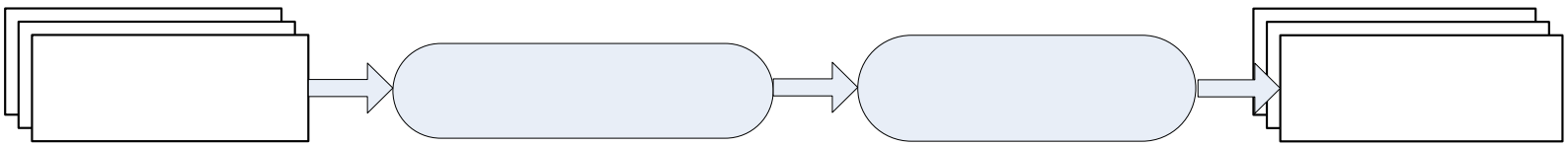
- В 1933 году изобретатель П.П.Смирнов-Троянский получил в СССР патент на механическую «машину для подбора и печатания слов при переводе с одного языка на другой» - «Лингвистический арифмометр» .
- Он предложил и автоматический двуязычный словарь, и схему кодирования межъязыковых грамматических соответствий; правда, только для «синтетического» языка эсперанто.

40-е годы – первые системы МП

- Теоретической основой начального периода работ по машинному переводу был взгляд на язык, как кодовую систему.
- В марте 1947 специалист по криптографии Уоррен Уивер в своем письме Норберту Винеру впервые поставил задачу машинного перевода, сравнив ее с задачей дешифровки.
- В 1949 г. он составил меморандум, в котором теоретически обосновал принципиальную возможность создания систем машинного перевода.

Концепция Interlingva

- Идеи Уивера легли в основу подхода к МП, основанного на **концепции Interlingva**: стадия передачи информации разделена на два этапа.
  - На первом этапе исходное предложение переводится на язык-посредник (созданный на базе упрощенного английского языка),
  - На втором этапе результат этого перевода представляется средствами выходного языка.



Первые системы МП

- В 1952 г. состоялась первая конференция по МП в Массачусетском технологическом университете.
- В 1954 г. **Джорджтаунский эксперимент**. В Нью-Йорке была представлена первая система МП — **IBM Mark II** (словарь в 250 единиц и 6 грамматических правил), осуществлявшая перевод с русского языка на английский.
- В 1954 г. первый эксперимент по МП был осуществлен в СССР И.К. Бельской и Д.Ю. Пановым в Институте точной механики и вычислительной техники АН СССР.
- Первая система МП с английского языка на русский на универсальной вычислительной машине была разработана коллективом под руководством Ю. А. Моторина.

Системы прямого перевода

- Причины невысокого качества МП в 50-е годы были:
  - ограниченные возможности аппаратных средств:
    - малый объем памяти;
    - низкая скорость доступа к информации;
  - невозможность полноценного использования языков программирования высокого уровня;
  - отсутствие теоретической базы по компьютерной лингвистики.
- Системы МП первого поколения – **системы прямого перевода (СПП)** – представляли собой программно-аппаратные комплексы, анализирующие текст пословно «слово за словом» (word-to-word) без синтаксической и смысловой целостности.

Системы МП в 60-е годы

Разработка систем МП в 60-е годы:

- в США при финансировании Мормонской церкви;
- в Канаде (например система METEO);
- в Европе — группами GENA (Гренобль) и SUSY (Саарбрюкен);
- в СССР (Москва) отечественными лингвистами (И.А. Мельчук и Ю.Д. Апресян) — лингвистический процессор ЭТАП.

Новый импульс в разработке систем МП (70-80-е годы)

- Новый подъем исследований в области МП был связан с серьезными достижениями в области искусственного интеллекта, а создание систем машинного перевода было осмыслено в 1970-е годы как одна из частных задач этого нового исследовательского направления.
- Исследователи ставили целью развитие «реалистических» систем МП, предполагавших участие человека на различных стадиях процесса перевода.

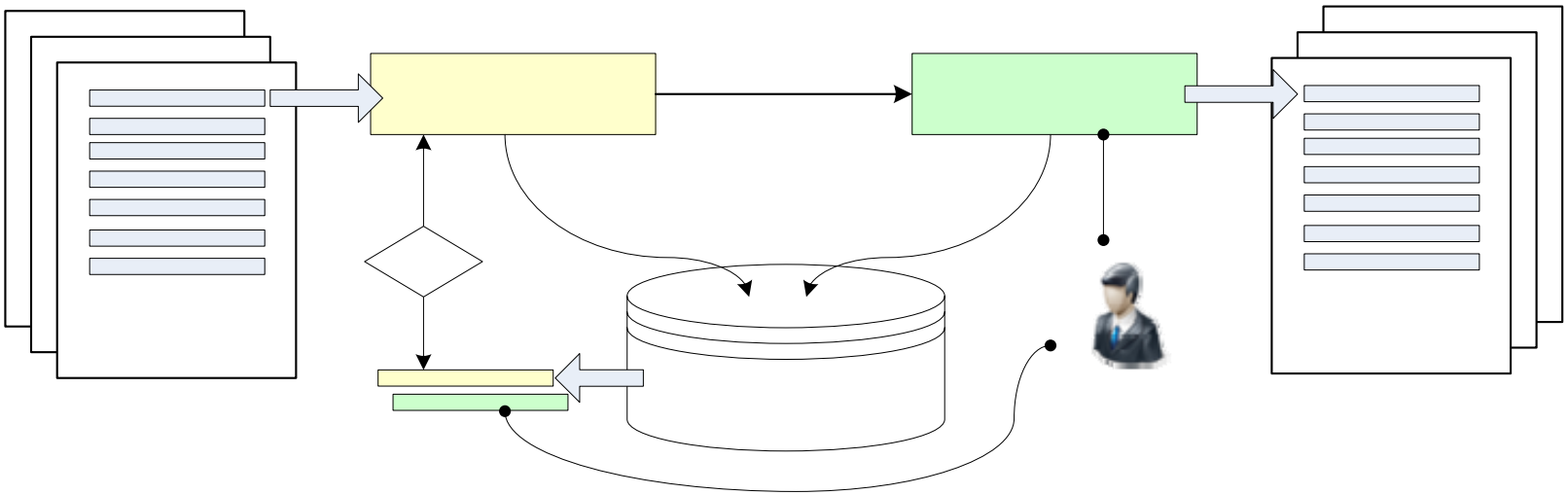
Стимулы к развитию МП

Можно выделить два стимула к развитию работ по МП:

- первый – **научный**, он определяется комплексностью и сложностью компьютерного моделирования перевода;
- второй – **социальный**, обусловлен он возрастающей ролью самой практики перевода в современном мире как необходимого условия обеспечения межъязыковой коммуникации, объем которой возрастает с каждым годом.

Технология TM (translation memory)

- В процессе перевода сохраняется исходный сегмент текста(предложение) и его перевод;
- если подобный исходному сегмент обнаруживается, он отображается вместе с переводом и указанием совпадения;
- затем переводчик принимает решение (редактировать, отклонить или принять перевод), результат которого сохраняется системой.



Советские системы МП 70-80 гг.

- В СССР с середины 70-х годов были созданы промышленные системы МП:
  - АМПАР (английский ➡ русский);
  - НЕРПА (немецкий ➡русский);
  - ФРАП (французский ➡ русский);
  - АСПЕРА (русский ➡ английский).
  - Автоматические терминологические словари.

- На этих разработках основываются такие системы МП, как Stylus, Socrat и другие.

Современные системы МП

- В настоящее время несколько десятков компаний занимаются разработкой коммерческих систем МП, в их числе: Systran, IBM, L&H (Lernout & Hauspie), Language Engineering Corporation, Transparent Language, Nova Incorporated, Trident Software, Atril, TRADOS, Caterpillar Co., LingoWare; Ata Software; Lingvistica b.v. и др.
- В июле 1990 года на выставке PC Forum в Москве была представлена первая в России коммерческая система машинного перевода под названием PROMT (PROgrammer's Machine Translation).

Недостатки современных систем МП

- Фактически всеми системами осуществляется перевод только на уровне поверхностного синтаксиса, поскольку еще не разработаны эффективные модели формального представления смысла.
- Установка на жанровую ограниченность текстов привела к тому, что задача моделирования ЕЯ фактически уступила место задаче моделирования ограниченных (и крайне примитивных) подъязыков отдельных отраслей знания.

Системы МП

- PROMT http://www.e-promt.ru/
- Pragma http://www.trident.com.ua/
- TRADOS http://www.trados.com/
- DEJA VU http://www.atril.com/

On-line сервисы МП

- Сервис перевода текстов Google Translate http://translate.google.com/
- Online Система машинного перевода Babel Fish (Systran) на портале Altavista http://babelfish.yahoo.com/
- Online-переводчик компании ПРОМТ http://www.translate.ru/
- Online Языковые инструменты комапании Google (перевод текста, сайтов) http://www.google.ru/language_tools
- Online-переводчик InterTran (перевод между 29 языками) http://www.tranexp.com:2000/Translate/result.shtml
- translation.langenberg.com (страница с вызовом online-переводчиков различных производителей и иных лингвистических программ) http://translation.langenberg.com/

Состав компонент систем МП

- двуязычные словари, снабженные необходимой информацией (морфологической, синтаксической и семантической);
- средства грамматического анализа, в основе которых лежит формальная грамматика.

Системы МП: этапы анализа

- 1 этап анализа.
- Осуществляется ввод текста и поиск входных словоформ во входном словаре 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, например, каким членом предложения может быть данное слово.

Независимый синтаксический анализ

- Машина осуществляет синтаксический анализ предложения без опоры на значения составляющих его слов, с использованием информации только об их грамматических свойствах.
- Пример:
  Глокая куздра штетко будланула бокра и кудрячит бокрёнка *(Л.В.Щерба).*

  «куздра» – это существительное-подлежащее,
  «глокая» – определение к существительному,
  «будланула» – глагол-сказуемое,
  «штетко» – обстоятельство образа действия,
  «бокра» – дополнение,
  «кудрячит» – глагол-сказуемое,
  «бокрёнка» – дополнение.

Синтаксический анализ

- В результате синтаксического анализа возникает синтаксическая структура, которая изображается в виде дерева зависимостей: «корень» – сказуемое, а «ветви» – синтаксические отношения его с зависимыми словами.
- Каждое слово предложения записывается в своей словарной форме, а при ней указываются те грамматические характеристики, которыми обладает это слово в анализируемом предложении.

Системы МП: этапы анализа

2-й этап включает в себя:

- перевод идиоматических словосочетаний, фразеологических единств или штампов данной предметной области;
- определение основных грамматических (морфологических, синтаксических, семантических и лексических) характеристик элементов входного текста;
- разрешение неоднозначности;
- анализ и перевод слов.

Системы МП: этапы анализа

- 3 этап - 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, притом, что в оригинале может быть и единственное число.

Системы МП: этапы анализа

- 4 этап. Синтез выходных словоформ и предложения в целом на выходном языке.
- В машинную память помимо наборов синтаксических правил для каждого языка «вкладывают» и правила преобразования синтаксических структур. К этому добавляют правила перехода от уже преобразованной структуры к предложению того языка, на который делается перевод. Такой переход от структуры к реальному предложению называется синтаксическим синтезом.

Системы МП: этапы анализа

- 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.

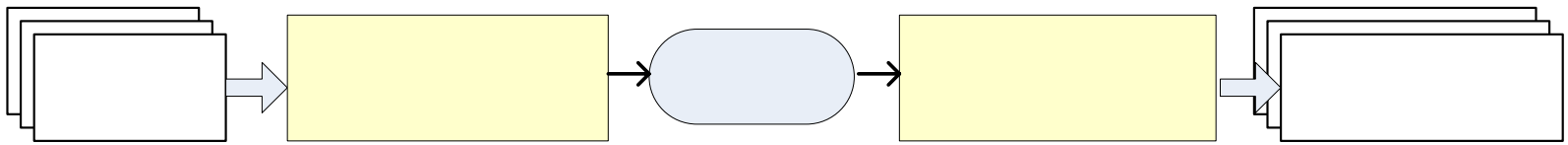
Проблема перевода текста

- Перевод текста с одного естественного языка (ЕЯ) на другой складывается из двух последовательных задач:
  - понимание текста
  - синтеза текста.

- «Язык есть универсальный преобразователь смысла в текст и обратно» [И.А.Мельчук]

Упрощенный алгоритм МП

- Основная идея алгоритма МП состоит в следующем: процесс перевода состоит из двух последовательных операций: анализа входного текста, т.е. отыскания его смысла, и синтеза выходного текста, т.е. построения по заданному смыслу текста на выходном языке.



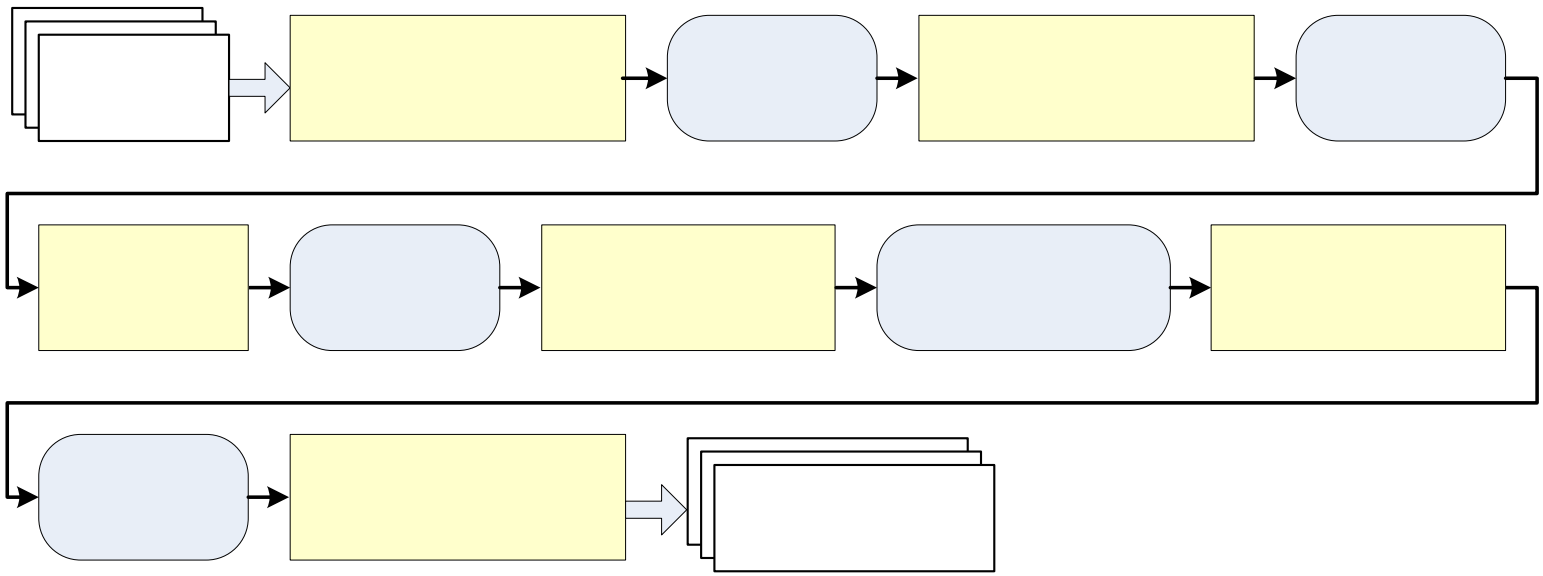
Семантическое представление текста

- Смысл, в отличие от текста, который произносится или пишется, - объект ненаблюдаемый, существующий лишь в человеческом мозгу. Поэтому зафиксировать (построить) в системе МП невозможно.
- В современной лингвистике используется понятие семантического представления (СемП) текста - некоторого конструкта, который условно принимается за смысл и может быть записан.

Уровень семантического представления

- При анализе текста сначала строится его морфологическое представление, по морфологическому представлению - синтаксическая структура, и только по синтаксической структуре строится семантическое представление.
- На практике в системах МП используют несколько упрощенный алгоритм: при анализе преобразование текста не доводится до уровня семантического представления, а останавливается на некотором промежуточном уровне — например, на уровне синтаксической структуры.

Алгоритм МП



Описание системы МП «Этап 3»

Морфологическая структура входного языка

- Морфологическая структура (МорфС) предложения - последовательность лемм и приписанных им морфологических характеристик. Например, предложение:

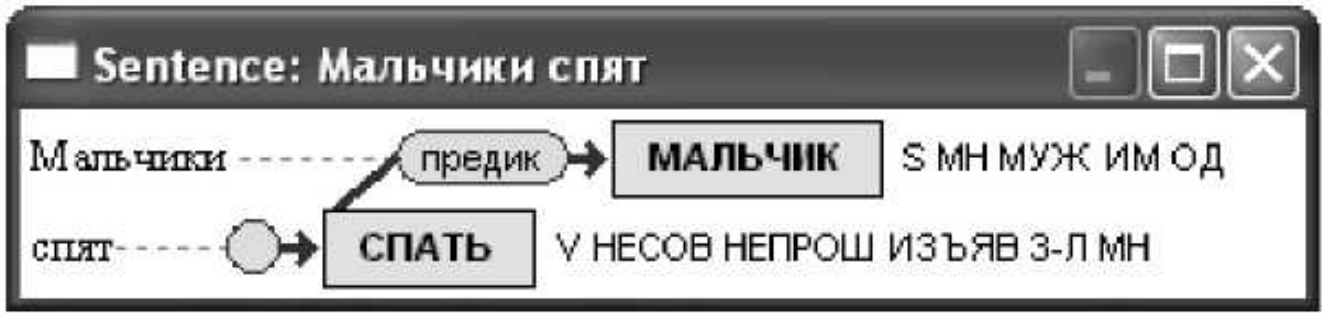
  Мальчики спят.

  МорфС предложения:

  МАЛЬЧИК s,мн,им **СПАТЬ** v,несов,непрош,изъяв,3-л,мн

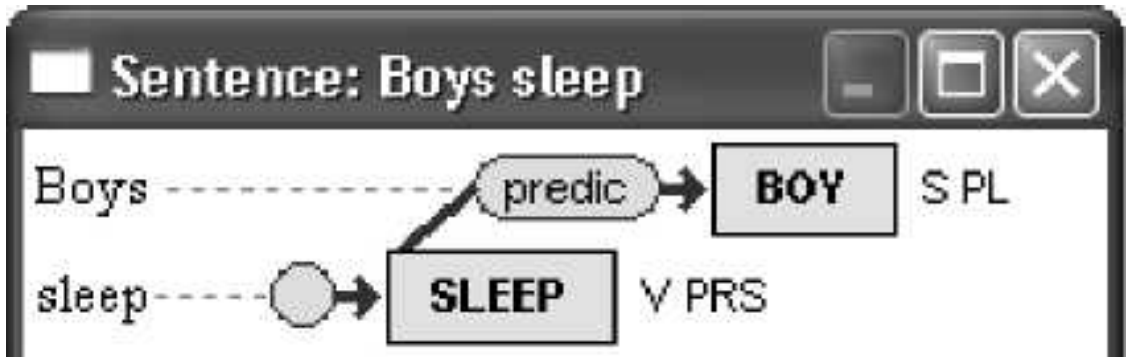
Синтаксическая структура входного языка

- Синтаксическая структура (СинтС) предложения представляется в виде дерева зависимостей - объекта, состоящего из лемм с приписанными им характеристиками (узлами СинтС), соединенными стрелками, которые помечены именами синтаксических отношений.



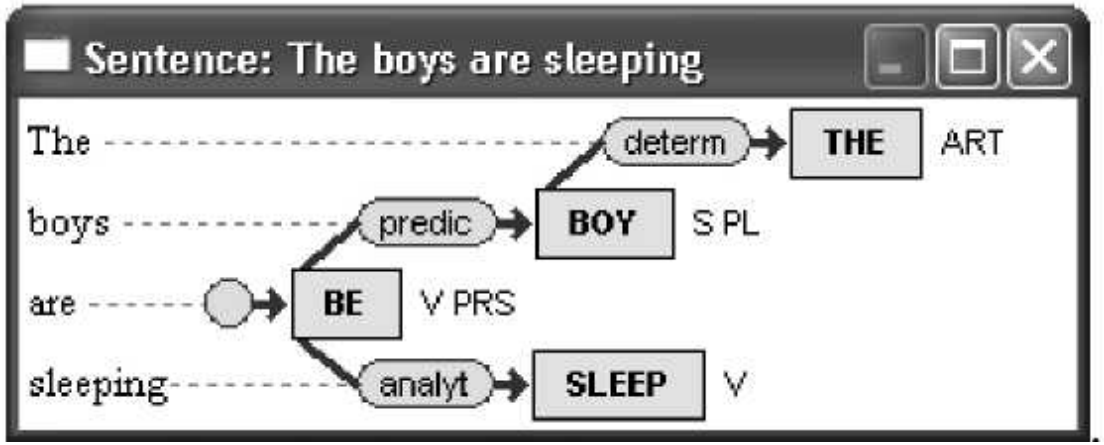
Перевод синтаксической структуры

- Блок перевода заменяет элементы входного языка (слова, характеристики, синтаксические отношения) элементами выходного языка, строит СинтС предложения на выходном языке.



Нормализация синтаксической структуры

- Выходная СинтС подвергнется нескольким промежуточным этапам преобразования, которые ее нормализуют: устанавливают правильный порядок слов, добавят при необходимости нужные узлы (например, артикли, сильноуправляемые предлоги, вспомогательные глаголы).



Морфологическая структура предложения на выходном языке

- Блок синтаксического синтеза приписывает элементам СинтС недостающие морфологические характеристики и строит МорфС выходного предложения.

  THE Art **BOYS** s,pl **BE** v,prs,pl **SLEEP** v,ing

- Далее блок морфологического синтеза построит предложение в обычном виде:

  The boys are sleeping.

Пример действий системы МП при переводе простого предложения

- Для перевода ***Я тебе нравлюсь?*** в ***Do you like me?***
- Основные действия системы МП:
  - 1) разрешить неоднозначность словоформы *тебе* (которая может быть дательным, а может быть предложным падежом лексемы ТЫ);
  - 2) построить СинтС предложения, для чего определить вершину этого предложения (сказуемое *нравлюсь),* его подлежащее *(я)* и дополнение *(тебе);*
  - 3) сравнить словарные статьи глагола*нравиться* и его переводного эквивалента *like,* убедиться, что при переводе подлежащее и дополнение меняются ролями — подлежащее *нравиться* становиться дополнением *like,* и наоборот; 4) — и осуществить соответствующую перестройку структуры;
  - 5) переставить дополнение к глаголу *like* - слово *me* - в постпозицию к этому глаголу;
  - 6) породить вспомогательный глагол *do* и поставить его на нужное место.

Языковая неоднозначность

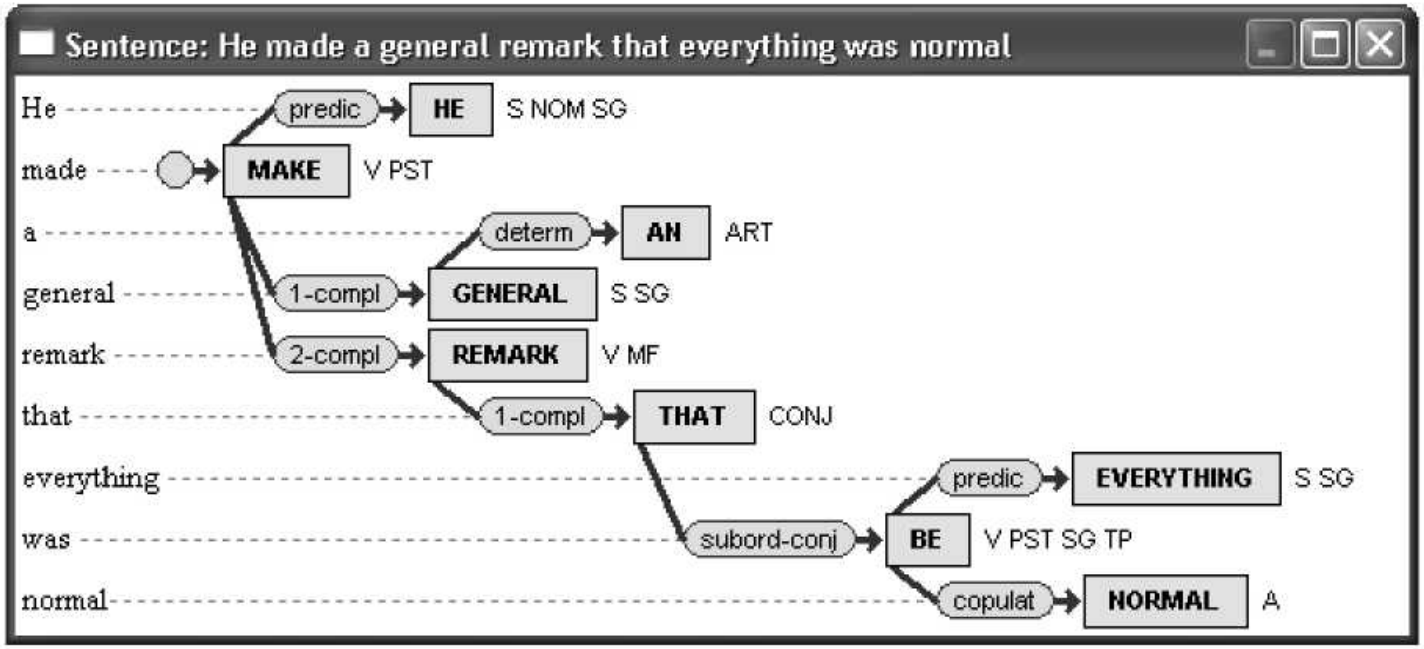
- Синтаксическая омонимия:
- Например, в предложении ***Я принес лук*** вне контекста невозможно понять, о каком луке — овоще или оружии — идет речь. В выражении ***Приглашение доктора*** равным образом нельзя понять, является ли доктор приглашенным или приглашающим, а в предложениях типа ***Мать любит дочь*** нельзя понять, кто кого любит.
- В этом случае система МП может предлагать варианты перевода.

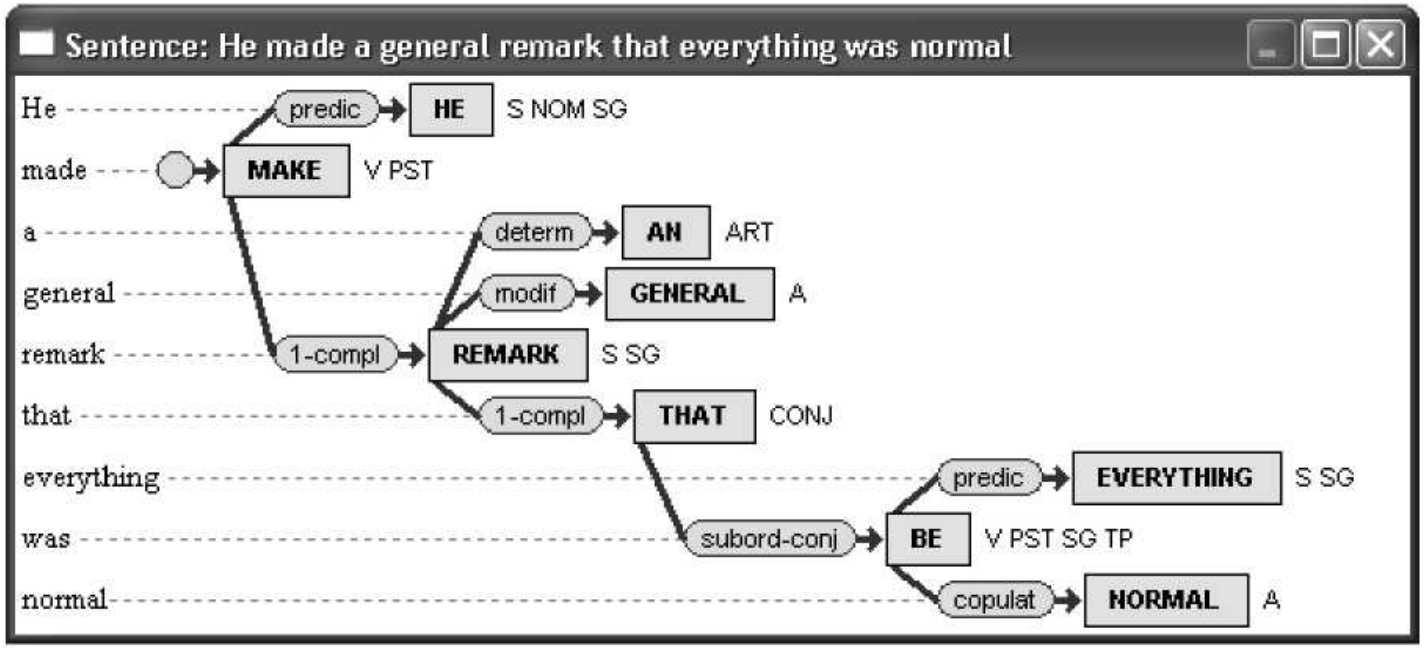
Проблема многозначности слов

- Для решения проблемы многозначности слова используется анализ контекста.
- У каждого из «конкурирующих» (при интерпретации) значений – свой набор контекстов. И именно вот эта зависимость значения от окружения позволяет слушающему понять высказывание правильно.
- Для правильного понимания высказывания необходимо в полной мере учитывать также правила обусловленности выбранного значения лексическим окружением, семантическим контекстом, грамматическим (морфолого-синтаксическим) контекстом.

Примеры различных вариантов перевода системы МП

- ***Моих родителей звали Иван и Мария.***
- Варианты перевода:
  - ***My parents' names were Ivan and Maria*** ('именами моих родителей были Иван и Мария');
  - ***It is Ivan and Maria that called my parents*** ('Иван и Мария приглашали куда-то моих родителей').
- ***He made a general remark that everything was normal.***
- Варианты перевода:
  - Он сделал общее замечание, что все нормально;
  - Он заставил генерала заметить, что все нормально.

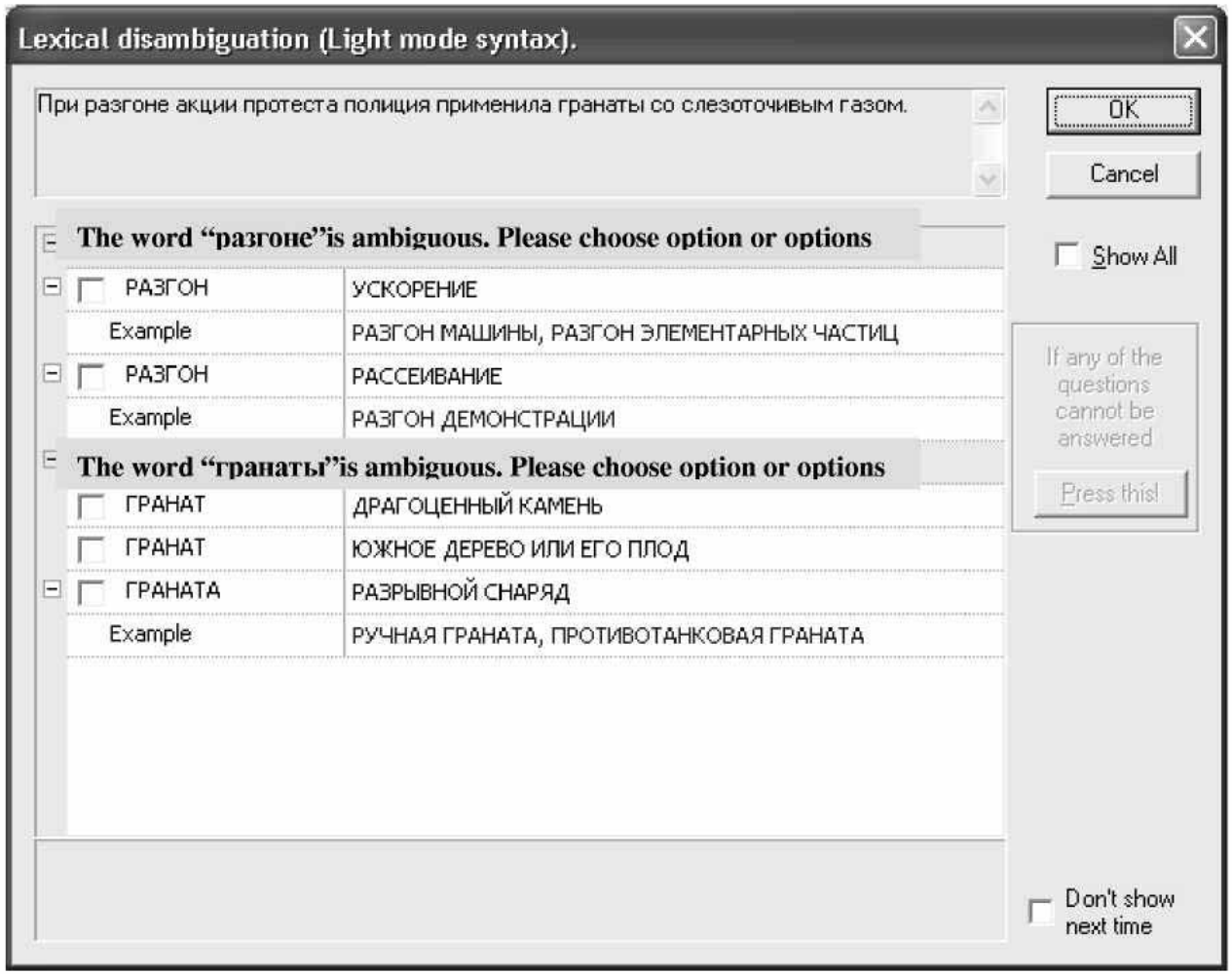

Sentence: He made a general remark that everything was normal
He
made
a
general
remark
that
everything
was
normal
predic
HE
S NOM SG
MAKE
V PST
determ
AN
ART
1-compl
GENERAL
S SG
2-compl
REMARK
V MF
1-compl
THAT
CONJ
predic
EVERYTHING
S SG
subord-conj
BE
V PST SG TP
copulat
NORMAL
A


Sentence: He made a general remark that everything was normal
He
made
a
general
remark
that
everything
was
normal
predic
HE
S NOM SG
MAKE
V PST
determ
AN
ART
modif
GENERAL
A
1-compl
REMARK
S SG
1-compl
THAT
CONJ
predic
EVERYTHING
S SG
subord-conj
BE
V PST SG TP
copulat
NORMAL
A

Примеры неправильного перевода

- ***Для разгона акции протеста полиция применила гранаты со слезоточивым газом***
- ***For acceleration of the shares, police applied garnets with tear gas.***
  - слово *разгон* было идентифицировано в значении 'ускорение' (как *в разгон автомобиля* или *разгон элементарных частиц),*
  - *акция* - 'ценная бумага' (как в *купили акции металлургического завода),*
  - а *гранаты* - как множественное число слова *гранат* 'драгоценный камень‘.

Модуль интерактивного разрешения неоднозначности



Качество систем МП

- Действующие системы машинного перевода, как правило, ориентированы на конкретные пары языков (например, французский и русский или японский и английский) и используют, как правило, переводные соответствия либо на поверхностном уровне, либо на некотором промежуточном уровне между входным и выходным языком.
- Качество машинного перевода зависит от объема словаря, объема информации, приписываемой лексическим единицам, от тщательности составления и проверки работы алгоритмов анализа и синтеза, от эффективности программного обеспечения.

Системы МП, подходы

- На практике переводческой деятельности и в информационной технологии различаются два основных подхода к машинному переводу.
- С одной стороны, результаты машинного перевода могут быть использованы для поверхностного ознакомления с содержанием документа на незнакомом языке (в этом случае текст не требует тщательного редактирования).
- Другой подход предполагает использование машинного перевода вместо обычного «человеческого» (это предполагает тщательное редактирование и настройку системы перевода на определенную предметную область).