УРОКИ МАШИННОГО ПЕРЕВОДА ДЛЯ ДЕТЕЙ И ВЗРОСЛЫХ[1]

Л. Л. Иомдин
Институт проблем передачи информации РАН
iomdin@cl.iitp.ru

Вводные замечания

Что такое машинный, он же автоматический, он же компьютерный, перевод? Сейчас, когда для перевода текстов с одного языка на другой компьютер используется самыми разными способами – от двуязычных и многоязычных электронных словарей до систем типа translation memory («память», или «архив» переводов), этот вопрос оказывается не таким уж простым. Мы будем понимать машинный перевод как процесс, при котором компьютер по заданному тексту на одном языке **производит новый текст на другом языке,** которого раньше в этом компьютере не было: понятно, что ни словари, ни архивы переводов таким свойством не обладают.

Когда можно говорить о том, что текст А на одном естественном языке является переводом текста Б на другом языке? Разумеется, тогда, когда оба текста – А и Б – имеют одинаковый смысл. Задача любого переводчика как раз и состоит в том, чтобы передать смысл текста (будь то письменного или устного) на одном языке средствами другого языка. В этом же состоит и задача машинного перевода.

Посмотрим, из каких частей состоит процесс перевода. Когда переводчик (или компьютер) получает некоторый подлежащий переводу текст, он, прежде всего, должен этот текст понять, т.е. постичь его смысл. Этот смысл надо теперь облечь в слова, т.е. создать новый текст. Таким образом, перевод складывается из двух последовательных задач: задачи понимания текста и задачи производства текста.

Обратим теперь внимание на то, что эти две задачи постоянно решают все люди, пользующиеся естественным языком, даже те, которые знают только свой родной язык и ни с какими переводами и переводчиками никогда не имели дела. Если человек хочет сообщить собеседнику свою мысль (не вдаваясь в подробности, укажем, что мысль и смысл очень близки), он порождает текст (письменный или устный), выражающий эту мысль. Этот текст может состоять из одного междометия, а может занимать несколько томов. Тем самым носитель естественного языка (говорящий, автор) решает вторую из сформулированных только что задач перевода – задачу производства, или, как говорят лингвисты, синтеза текста. Другой носитель языка (слушающий, читатель), получив этот текст, решает первую из вышеупомянутых задач перевода – задачу понимания, или анализа текста. Поскольку каждый человек при общении с другими людьми попеременно оказывается то говорящим, то слушающим, ему и приходится решать как задачу понимания, так и задачу производства текста.

Именно эти две задачи составляют основу лингвистики – науки, предметом которой является человеческий язык. Недаром знаменитый лингвист И.А.Мельчук говорил, что язык есть универсальный преобразователь смысла в текст и обратно, а его лингвистическая теория так и называется: теория «Смысл ⇔ Текст». Достижения современной

[1] Данная работа выполнена при частичной поддержке гранта Российского фонда фундаментальных исследований 02-06-80085. Автор выражает Фонду искреннюю признательность.

лингвистической науки в изучении процессов анализа и синтеза естественно-языкового текста, в моделировании этих процессов очень велики. Не удивительно поэтому, что эти достижения стремятся как можно полнее использовать разработчики систем машинного перевода. В процессе анализа *входного языка* системы машинного перевода (МП) используют средства, предлагаемые лингвистикой для моделирования механизмов понимания, а в процессе синтеза *выходного языка* – средства, наработанные лингвистикой для моделирования механизмов производства текста. Одна из таких систем – система ЭТАП-3 [1-2], в создании которой активно участвует автор этих строк, – больше других опирается на лингвистическую теорию «Смысл ⇔ Текст».

Как именно работает эта система – детище Института проблем передачи информации РАН, мы сейчас увидим.

Алгоритм машинного перевода

Наша система работает с письменными текстами. Основными рабочими языками системы являются русский и английский: ЭТАП-3 осуществляет перевод между этими языками в обоих направлениях. Кроме того, в ЭТАП-3 входит несколько экспериментальных прототипов систем французско-русского, русско-испанского, русско-немецкого, русско-корейского и арабско-английского МП – об этих прототипах мы сейчас говорить не будем.

Основная идея алгоритма МП состоит в следующем. Как мы уже видели, 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. Смысл при этом является единым для входного и выходного текстов, или, как говорят лингвисты, *инвариантом* для них обоих. Тем самым идеальным решением было бы построить систему из двух блоков. Первый из этих блоков, получая на входе текст А, конструировал бы для него смысл, а второй, получая на входе этот смысл, конструировал бы по нему текст Б.

Эта идеальная схема наталкивается, однако, на значительные трудности. Заметим, прежде всего, что смысл, в отличие от текста, который произносится или пишется, – объект ненаблюдаемый, существующий лишь в человеческом мозгу. «Построить его» непосредственно нам не удастся ни при каких обстоятельствах. Для преодоления этой трудности в современной лингвистике было выработано понятие *семантического представления* (СемП) текста – некоторого конструкта, который условно принимается за смысл и который, в отличие от смысла, можно записать (на бумаге или на экране компьютера). СемП считается независимым от конкретного языка, универсальным объектом. Соответственно, идеальный алгоритм МП заменяется более реальным, а именно: на первом этапе по входному тексту строится СемП, а на втором по этому СемП строится выходной текст.

Весь опыт лингвистики свидетельствует, однако, о том, что переход от текста непосредственно, в один шаг, к семантическому представлению – задача невыполнимая. Равным образом невыполнима и обратная задача – непосредственный переход от СемП к тексту. Для решения этих задач лингвистика выработала механизмы лингвистических *уровней* – переход от текста к СемП и обратно осуществляется не в один шаг, а в несколько шагов. Например, при анализе текста сначала строится его морфологическое представление, по морфологическому представлению – синтаксическая структура, и только по синтаксической структуре строится СемП. Аналогичным образом на несколько шагов распадается и задача синтеза текста.

Далее, выясняется, что и эта схема оказывается трудновыполнимой – в первую очередь потому, что универсальное семантическое представление – объект невероятной сложности.

Поэтому на практике в системах МП используют несколько упрощенный алгоритм, а именно: при анализе преобразование текста не доводится до уровня СемП, а останавливается на некотором промежуточном уровне – например, на уровне синтаксической структуры. Преобразование синтаксической структуры входного языка в синтаксическую структуру выходного языка осуществляется с помощью специально созданных правил. Именно так поступает и система ЭТАП-3. Рассмотрим алгоритм ее работы несколько подробнее.

На вход системы поступает предложение на входном языке. Пусть, к примеру, работает русско-английская версия системы ЭТАП-3. На вход системы поступает предложение на входном (русском) языке. Это предложение подвергается сначала морфологическому анализу, в результате которого появляется морфологическая структура (МорфС) предложения – последовательность лемм (слов в словарной форме) и приписанных ей морфологических характеристик. Например, МорфС предложения (1) *Мальчики спят* будет выглядеть приблизительно так:

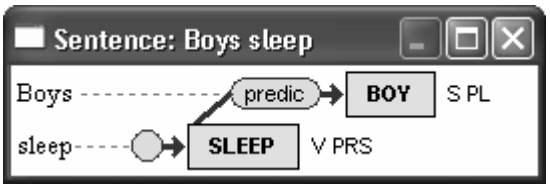
МАЛЬЧИК$_{S,мн,им}$ СПАТЬ$_{V,несов,непрош,изъяв,3-л,мн}$

Сокращения в подстрочных индексах при лемме МАЛЬЧИК означают здесь существительное, именительный падеж, множественное число, а при лемме СПАТЬ – глагол, несовершенный вид, непрошедшее (в данном случае настоящее) время, изъявительное наклонение, третье лицо, множественное число.

МорфС предложения подвергается синтаксическому анализу, в результате которого появляется синтаксическая структура (СинтС) предложения. В системе ЭТАП-3 используется СинтС предложения в виде так называемого дерева зависимостей – объекта, состоящего из лемм с приписанными им характеристиками (узлами СинтС), соединенными стрелками, которые помечены именами синтаксических отношений. Например, СинтС предложения (1) имеет вид[2]

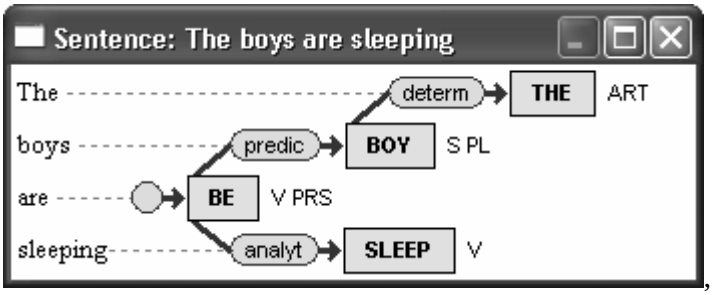
Единственная стрелка дерева зависимостей – предик(ативная) – характеризует связь между сказуемым *спят* в качестве главного члена предложения и его подлежащим *мальчики* в качестве зависимого члена.

Синтаксическая структура входного предложения поступает на блок собственно перевода, который, заменяя элементы входного языка (слова, характеристики, синтаксические отношения) элементами выходного языка, строит СинтС предложения на выходном языке. Для (1) эта СинтС будет выглядеть так:



[2] Здесь и далее воспроизводятся реальные структуры, генерируемые системой ЭТАП-3.

Далее выходная СинтС подвергнется нескольким промежуточным этапам преобразования, которые ее нормализуют – линеаризуют (т.е. установят правильный порядок слов), добавят при необходимости нужные узлы (например, артикли, сильноуправляемые предлоги, вспомогательные глаголы). Нормализованная СинтС, которая (если мы решим использовать в переводе время Present Continuous) для предложения (1) примет вид[3]



поступает на вход блока синтаксического синтеза, который при необходимости припишет элементам СинтС недостающие морфологические характеристики и построит МорфС выходного предложения. Эта МорфС для (1) будет выглядеть так:

$THE_{Art}\ BOY_{S,pl}\ BE_{V,prs,\ pl}\ SLEEP_{V,ing}$

Внимательный читатель заметит, что сравнительно с нормализованной СинтС в МорфС появились две новые характеристики глаголов: глагол BE приобрел характеристику pl (множественное число, сформированное правилом согласования с подлежащим), а глагол SLEEP = характеристику ing (причастие настоящего времени, используемое в Present Continuous).

Наконец, МорфС выходного предложения поступит на вход последнего блока системы – морфологического синтеза, который построит для нее предложение в обычном орфографическом виде: *The boys are sleeping.*

Не будет большим преувеличением сказать, что приблизительно так работает большинство существующих систем машинного перевода. Самым важным общим свойством этих систем является, пожалуй, разграничение грамматики языка (т.е. правил, описывающих разные уровни языка) и словаря.

Машинный перевод – трудное дело

Разумеется, если бы все предложения естественного языка были такими простыми, как предложение (1), идеальная система машинного перевода давным-давно была бы построена. К сожалению, дело обстоит куда сложнее. Чтобы лучше оценить проблемы, с которыми сталкиваются разработчики систем МП, попытаемся присмотреться внимательнее к перечню задач, которые приходится решать такой системе при обработке даже самых простых предложений. Сравним русское вопросительное предложение (2) *Я тебе нравлюсь?* и его английский перевод *Do you like me?* и перечислим хотя бы основные действия, которые должна выполнить система МП, чтобы обеспечить получение этого перевода. Итак, 1) необходимо разрешить неоднозначность словоформы *тебе* (которая может быть дательным, а может быть предложным падежом лексемы ТЫ); 2) построить СинтС предложения (2), для чего определить вершину этого предложения (сказуемое *нравлюсь*), его подлежащее (*я*) и дополнение (*тебе*); 3) сравнить словарные статьи глагола

[3] В этой нормализованной СинтС, как нетрудно увидеть, появились два новых узла – артикль при существительном *boys* и вспомогательный глагол *be*, необходимый для того, чтобы построить время Present Continuous.

нравиться и его переводного эквивалента *like*, убедиться, что при переводе подлежащее и дополнение меняются ролями – подлежащее *нравиться* становиться дополнением *like*, и наоборот; – 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.

Список этих операций представляется достаточно впечатляющим, если учесть, что предложение (2) состоит из трех простых слов. Тот факт, что в деле создания систем МП за последнее десятилетие был достигнут значительный прогресс, не может не вызывать у создателей таких систем чувства гордости за проделанную работу.

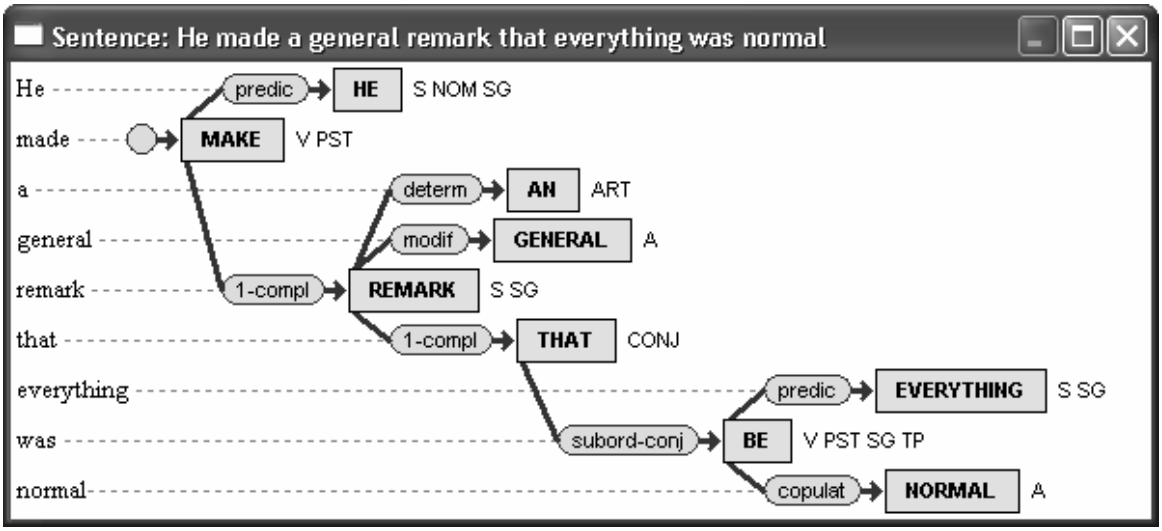
Языковая неоднозначность

Опыт показывает, что самая сложная проблема в машинном переводе – как, впрочем, и в любой другой системе автоматической обработки текстов на естественном языке – это проблема разрешения языковой неоднозначности.

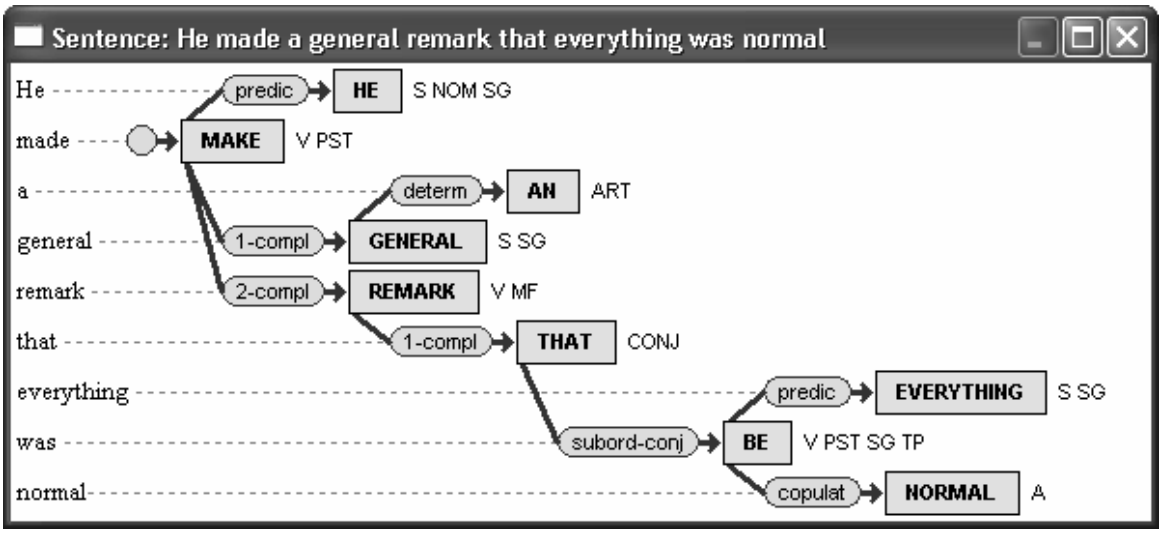
С простейшим примером такой неоднозначности – морфологической – мы уже познакомились в предложении (2), когда заметили, что словоформа *тебе* может быть как дательным, так и предложным падежом слова *ты*. Однако с ситуациями языковой неоднозначности лингвистам приходится сталкиваться на каждом шагу, и далеко не все они столь безобидны. Предложение (2) содержит неоднозначную словоформу *тебе*, но само по себе является совершенно однозначным: ни при каком прочтении этого предложения словоформа *тебе* не может проявить себя как предложный падеж. Благодаря этому и перевод (2) получился правильным и единственно возможным. Однако что делать системе, когда неоднозначность каких-то элементов переводимого предложения приводит к реальной двусмысленности всего предложения? 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* или *Весло повредило орудие* нельзя понять, кто кого любит и что чему нанесло повреждение (во всех подобных примерах принято говорить о синтаксической омонимии предложения). Нам представляется, что система МП в подобных случаях должна уметь давать все адекватные переводы – именно так и поступает ЭТАП-3, снабженная модулем множественного анализа.

В некоторых ситуациях система МП способна обнаружить лексическую неоднозначность и/или синтаксическую омонимию переводимого текста там, где она с трудом заметна человеку-переводчику. Например, ЭТАП-3 для предложения (3) *Моих родителей звали Иван и Мария* построил не только перевод *My parents' names were Ivan and Maria* (т.е. 'именами моих родителей были Иван и Мария'), но и перевод *It is Ivan and Maria that called my parents* ('Иван и Мария приглашали куда-то моих родителей'): дело, по-видимому, в том, что глагол ЗВАТЬ1 = 'называть по имени' у человека в присутствии в ближайшем контексте конкретных имен настолько прочно ассоциируется с ситуацией называния, что начисто вытесняет из памяти второе лексическое значение этого глагола ЗВАТЬ2 = 'приглашать прийти'. Для системы МП, у которой никаких ассоциаций, разумеется, нет, обе интерпретации предложения (3) равновероятны.

Аналогичным образом англо-русская версия системы ЭТАП-3 порождает для предложения типа (4) *He made a general remark that everything was normal* два равноправных перевода: 'Он сделал общее замечание, что все нормально' и 'Он заставил генерала заметить, что все нормально'. Разумеется, эти переводы получаются потому, что предложение (4) лексически и синтаксически неоднозначно. В первом случае СинтС предложения выглядит следующим образом:



Здесь слово *general* ‘общий’ является прилагательным-определением к существительному *remark* ‘замечание’, которое выступает в качестве прямого дополнения к глаголу *make* ‘делать’. Во втором же случае анализирующий компонент системы построил совершенно другую структуру. Слово *general* ‘генерал’ здесь оказывается существительным и выступает как первое дополнение к глаголу *make*, а слово remark ‘замечать’ – глагол, выступающий в роли второго дополнения при *make* и формирующий известную синтаксическую конструкцию типа Complex Object.



Рассмотрим теперь несколько менее очевидную ситуацию, с которой столкнулись разработчики системы ЭТАП-3. На первый взгляд, предложение (5) *Трое вышли из лесу* ничуть не сложнее по своему устройству, чем предложение (2) и не таит в себе никакой неоднозначности. Тем не менее в процессе экспериментальной эксплуатации системы это предложение получило весьма неожиданный перевод: *To Troy send from forest*. Внимательный читатель, конечно, сразу же понял, что здесь произошло: первое слово предложения (5) было идентифицировано системой не как числительное, а как дательный падеж существительного *Троя* (заметим, что это слово начинает предложение и поэтому пишется с заглавной буквы), а второе его слово – не как глагол *выходить* в прошедшем времени, а как глагол *высылать* в повелительном наклонении. Конечно, интерпретация предложения (5) как просьбы (дескать, когда будешь в лесу, отправь оттуда что-нибудь

городу Трое, пока он не погрузился под воду) представляется малоубедительной – но ведь она не исключена!

В чем состоит поучительность такого результата для разработчиков системы МП? В том, что в систему должны быть включены механизмы, позволяющие ей получать сначала более вероятные разборы входного текста (а, стало быть, и более вероятные его переводы).

Границы возможного

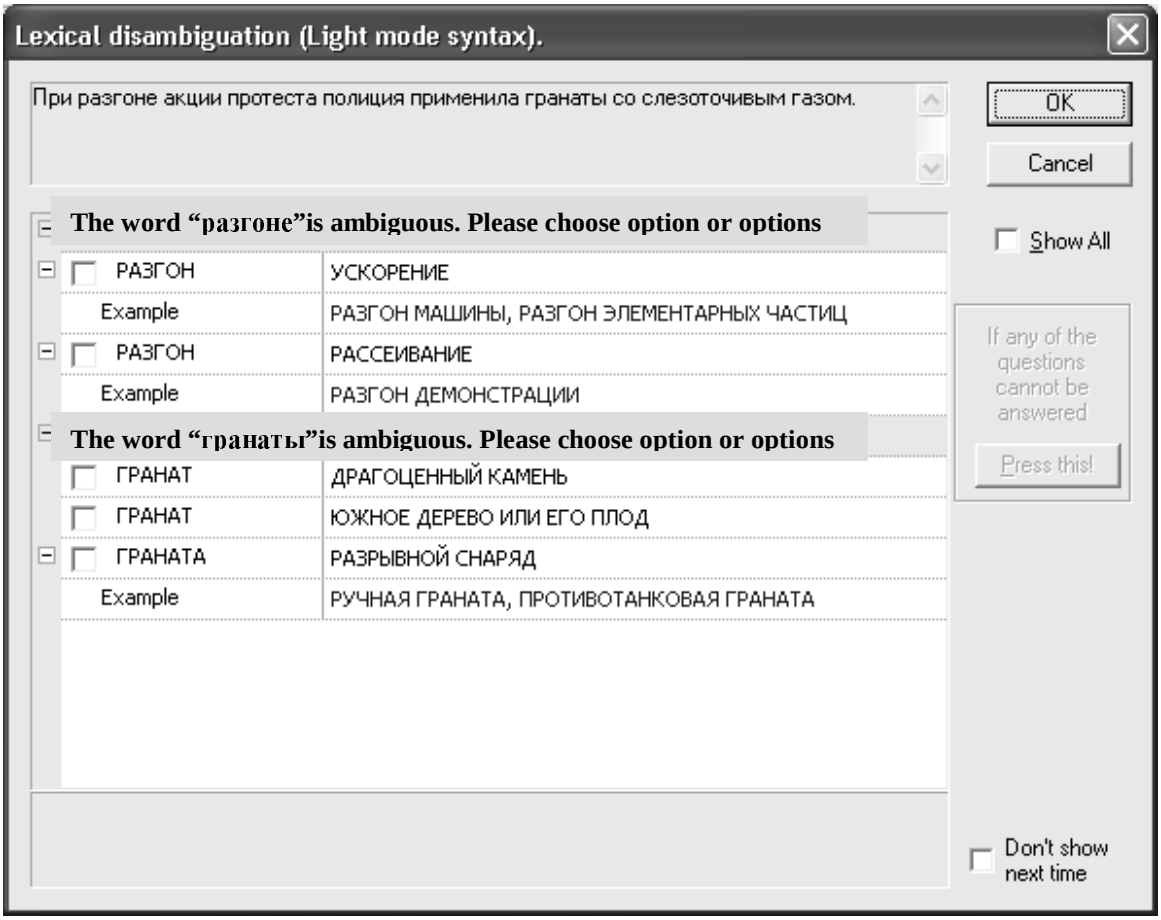
Однако даже не это главное при работе системы МП с текстами, содержащими неоднозначные элементы. Самые большие трудности при машинном переводе вызывают ситуации, когда неоднозначные элементы (скажем, лексические единицы) на самом деле не порождают реальной многозначности переводимого текста просто потому, что в условиях конкретного контекста некоторые интерпретации этих элементов оказываются невозможными.

Рассмотрим показательный пример из практики работы системы ЭТАП-3. Фраза (6) *Для разгона акции протеста полиция применила гранаты со слезоточивым газом* является, по нашему мнению, абсолютно однозначной. Тем не менее система МП выдала для нее английский перевод, из которого следовало, что сразу три существительных: *разгон*, *акция* и *гранаты* были проинтерпретированы неверно: слово *разгон* было идентифицировано в значении ‘ускорение’ (как в *разгон автомобиля* или *разгон элементарных частиц*), *акция* – как ‘ценная бумага’ (как в *купили акции металлургического завода*), а *гранаты* – как множественное число слова *гранат* ‘драгоценный камень’ (даже не ‘фрукт’!). Нечего и говорить, что результат перевода оказался фантасмагорическим: *For acceleration of the shares, police applied garnets with tear gas*.

Каким образом можно избежать подобных результатов? К сожалению, общего ответа на этот вопрос не существует. Дело в том, что с какой бы степенью подробности лингвисты ни описывали условия встречаемости лексических единиц в тексте, сочетаемостные, семантические и лексические ограничения на их функционирование, все равно достичь стопроцентной точности нам не удастся. Какую информацию, например, можно было бы записать в словарную статью существительного *разгон* ‘ускорение’ (или соответствующего глагола *разгонять* ‘ускорять’)? Можно, конечно, указать, что в качестве дополнения при нем могут выступать слова со значением предмета. Однако разве не является предметом акция – ценная бумага? Далее, можно было бы, например, попытаться воспользоваться тем фактом, что слова *разгон* и *разгонять* в значении ‘рассредоточение, рассредоточивать’ допускают лишь слова со значением массового объекта, что могло бы повысить шансы установления связи между соответствующим значением существительного *разгон* и соответствующего значения слова *акция* ‘поступок, выступление’. Заметим, однако, что в значение самого слова *акция* (в отличие, скажем, от *митинга* или *сборища*) не входит смысл ‘массовый объект’: по сути дела, идея массовости акции передается именно с помощью таких слов, как *разгон* или *разгонять*. Наконец, что, кроме знания окружающей нас действительности, может помочь нам исключить существование как драгоценных камней, так и фруктов со слезоточивым газом? Однако систем машинного перевода, так же хорошо знающих окружающую действительность, как знает ее человек, пока не существует.

В значительной степени решить указанную проблему поможет разрабатываемый сейчас авторами системы ЭТАП-3 модуль интерактивного разрешения неоднозначности. В ситуациях, подобных (6), система МП в строго определенный момент работы алгоритма анализа текста обращается к человеку за помощью и просит его указать, какая именно интерпретация неоднозначного элемента текста имеется в виду.

Разумеется, человек, в отличие от компьютера, легко распознает нужные значения слов *разгон*, *акция*, *гранаты* и сообщит их системе. Вот как выглядит такое обращение на экране компьютера:



После того, как человек выберет нужные интерпретации, система исключит из рассмотрения все противоречащие им варианты и построит правильный перевод.

Приведем в заключение еще несколько поучительных примеров, когда система МП плохо справляется с разрешением неоднозначности текста: все они взяты из практики работы системы ЭТАП-3 (подробнее см. [3-4]).

Предложение (7) *На месте пожара был обнаружен мертвым пожилой сторож* любым носителем русского языка будет воспринято однозначно в смысле ‘На месте пожара кто-то обнаружил пожилого сторожа. Сторож был мертв’. Между тем анализирующему компоненту системы ЭТАП-3 ничто не помешало интерпретировать слово *мертвым* как деятеля при страдательном причастии *обнаружен* и в результате перевести (7) как *An elderly guard was found on the site of fire by the dead man.* На наш взгляд, ни в каком словаре и ни в каком правиле невозможно указать, что мертвец не может быть субъектом действия – потому что он в принципе может им быть! Правда, обнаружить кого-либо он вряд ли может, но вполне может кого-нибудь испугать или удивить. Однако данное знание слишком специфично, чтобы фиксировать его в формализованных языковых описаниях.

Предложение (8) *Было подтоплено 820 домовладений* (в предшествующем тексте речь шла о наводнении) было переведено как *820 households were heated* (т.е. ‘домовладения были подогреты’). Исключить такой перевод можно было бы разве что за счет формализации знания о том, что агентство новостей вряд ли мог бы заинтересовать факт, что кто-то немного подтопил печки в домах, пусть даже их было целых 820.

Предложение (9) *В 1999 году в ФРГ переехало 95 тысяч этнических немцев,* взятое с новостной ленты ИТАР-ТАСС, получило (наряду с совершенно правильным) перевод (9а) *In 1999 in the Federal Republic of Germany 95 thousand ethnic Germans were run over* (т.е. все эти люди были задавлены автомобилями). Как ни невероятно это звучит, интерпретация (9а) абсолютно законна, поскольку все мыслимые лексические и синтаксические условия для нее полностью соблюдены. Человек, конечно, не примет такой интерпретации и скажет, во-первых, что мир не так плох, чтобы столько народу погибло под колесами за один год, во-вторых, что погибших не стали бы характеризовать по национальной принадлежности, наконец, что слово «переехать» слишком разговорно, чтобы его применяло солидное агентство новостей. Но попробуйте все это формализовать!

Наконец, для предложения (10) *В этом году в России будет построено или отремонтировано 150 зданий судов* первый из полученных переводов звучал так: (10а) *In this year 150 buildings of ships in Russia will be erected or repaired.* Разумеется, зданий кораблей, в отличие от зданий трибуналов, на свете не бывает. Может быть, этот факт и можно записать в словарной статье слова *здание* в виде семантических ограничений на его зависимые – но только постфактум: составитель массового словаря не в состоянии предусмотреть все случайные факторы, которые могут потребовать таких ограничений. Ведь в нормальных текстах никому в голову не придет писать о зданиях чего бы то ни было, кроме учреждений, а осмысление, подобное (10а) – это результат случайных языковых «возмущений», вызванных «блуждающей» неоднозначностью.

Из сказанного можно сделать следующий важный вывод. Возможности машинного перевода ограничены принципиально. Среди многих причин ограниченности особо выделяется одна – невозможность эксплицировать достаточно простыми и формальными средствами все языковые факты, имеющие отношение к неоднозначности той или иной природы. Что ж поделаешь? Исследователи в любой науке вынуждены работать в условиях агрессивной среды, блуждающих токов и разнообразных возмущений. Будем продолжать работать в таких условиях и мы.

Литература

1. Ю. Д. Апресян, И. М. Богуславский, Л. Л. Иомдин, А. В. Лазурский, Н. В. Перцов, В. З. Санников, Л. Л. Цинман. Лингвистическое обеспечение системы ЭТАП-2. М.: Наука, 1989.

2. Jurij Apresian, Igor Boguslavsky, Leonid Iomdin, Alexander Lazursky, Vladimir Sannikov, Victor Sizov, Leonid Tsinman. ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT // *MTT 2003, First International Conference on Meaning – Text Theory*. Paris, École Normale Supérieure, Paris, June 16–18 2003, pp. 279–288.

3. Л. Л. Иомдин. Уроки русско-английского (из опыта работы системы машинного перевода) // *Труды Международного семинара Диалог'2002 по компьютерной лингвистике и ее приложениям.* Москва, Наука, 2002, Т. 2. С. 234–244.

4. Leonid Iomdin. Natural Language Processing as a Source of Linguistic Knowledge // *Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications.* Las Vegas, June 23–26 2003, pp. 68–74.