ИПМ им.М.В.Келдыша РАН • Электронная библиотека

Препринты ИПМ • Препринт № 89 за 2007 г.

ISSN 2071-2898 (Print)
ISSN 2071-2901 (Online)

В. А. Галактионов, А.М. Мусатов,
О.Ю. Мансурова, С.В. Ёлкин,
Э.С. Клышинский, В. Ю. Максимов,
С.Н. Аминева, Р.В. Жирнов,
С.Ю. Игашов, Т. Н. Мусаева

Система машинного перевода «Кросслятор 2.0» и анализ ее функциональности для задачи трансляции знаний

Рекомендуемая форма библиографической ссылки: Система машинного перевода «Кросслятор 2.0» и анализ ее функциональности для задачи трансляции знаний / В. А. Галактионов [и др.] // Препринты ИПМ им. М.В.Келдыша. 2007. № 89. 28 с.
https://library.keldysh.ru/preprint.asp?id=2007-89

О р д е н а Л е н и н а
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ
имени М.В. Келдыша
Р о с с и й с к о й а к а д е м и и н а у к

Галактионов В.А., Мусатов А.М.,
Ёлкин С.В., Клышинский Э.С.,
Максимов В.Ю.,
Аминева С.Н., Мансурова О.Ю.,
Мусаева Т.Н.

Система машинного перевода «Кросслятор 2.0» и анализ ее функциональности для задачи трансляции знаний

Препринт №

Москва

РОССИЙСКАЯ АКАДЕМИЯ НАУК
ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ имени М.В. Келдыша

Галактионов В.А., Мусатов А.М., Мансурова О.Ю., Ёлкин С.В., Клышинский Э.С., Максимов В.Ю., Аминева С.Н., Мусаева Т.Н.

Система машинного перевода «Кросслятор 2.0» и анализ ее функциональности для задачи трансляции знаний

Москва, 2007

Галактионов В.А., Мусатов А.М., Мансурова О.Ю., Ёлкин С.В., Клышинский Э.С., Максимов В.Ю., Аминева С.Н., Жирнов Р.В., Игашов С.Ю., Мусаева Т.Н.

Система машинного перевода «Кросслятор 2.0» и анализ ее функциональности для задачи трансляции знаний

Аннотация
В работе проведены обзор и классификация существующих систем машинного перевода, разобран состав и назначение логических блоков таких систем, приведена краткая история развития машинного перевода. На основе приведенной классификации рассматривается место системы машинного перевода «Кросслятор 2.0» среди современных систем автоматической обработки текстов. В соответствии с приведенной структурой систем машинного перевода дается подробный разбор функциональности системы «Кросслятор 2.0».
Итогом работы является анализ применимости системы «Кросслятор 2.0» для экспериментов в области трансляции знаний из одной предметной области в другую.

Работа поддержана РФФИ, грант 06-01-00538

Galaktionov V.A., Musatov A.M., Mansurova O.Yu., Yolkin S.V., Klyshinsky E.S. , Maximov V.Yu., Amineva S.N., Zhirnov R.V., Igashov C.Yu., Musaeva T.N.

The machine translation system “Crocclator 2.0” and its functionality analysis for the knowledge translation purposes

The paper presents the review and classification of existing Machine Translation Systems, discusses the structure and functions of logical blocks of such systems and shows the brief history of machine translation. On the basis of given classification the ranking of the Machine Translation System "Crosslator 2.0" among other modern automatic text processing systems is discussed. The detailed analysis of Crosslator 2.0 functionality has been carried out in accordance with the mentioned structure system.
The work is resulted in the analysis of "Crosslator 2.0" applicability for carrying out the experiments in the field of translation of knowledge from one subject domain to another.

Содержание

1. Введение

Задача трансляции знаний из одной предметной области в другую не может быть решена без использования методов машинной лингвистики. Дело в том, что перед тем как «понять» о чем идет речь в тексте и какой предметной области он относится, вначале необходимо понять его структуру, связи слов в предложении, то есть провести грамматический анализ текста. С другой стороны, трансляция текста производится во внутреннем представлении программной системы, осуществляющей такую трансляцию. В связи с этим после трансляции текста из одной предметной области в другую необходимо вновь просинтезировать текст из внутреннего представления в привычное человеку линейное.

Подобный перевод из текстового во внутреннее представление, а после обратно, характерен для систем машинного перевода. На данный момент сотрудники лаборатории АСОЛИ Института прикладной математики им. М.В. Келдыша накопили богатый опыт по обработке машинных текстов. Нами была разработана система машинного перевода «Кросслятор 2.0», работающая с русским, английским, испанским и турецким языками.

В данной работе мы попытаемся проанализировать, какая часть из имеющегося задела в области машинного перевода может быть использована в задаче трансляции знаний между предметными областями.

Вначале попытаемся дать характеристику созданной системы.

2. Обзор и классификация систем машинного перевода

Начиная с 1950-х годов в нашей стране и за рубежом интенсивно развиваются методы автоматической обработки текста (АОТ). До сих пор они остаются в центре внимания многих исследователей в области искусственного интеллекта.

Необходимость систем АОТ обусловливается большим ростом объема электронной текстовой информации, из которой нужно извлекать знания по возможности полностью автономными процедурами. Но развитие МП - интенсивно развивающейся области научных исследований и разработок СМП, в которых к процессу перевода с одного естественного языка на другой привлекается ЭВМ - определяется не только потребностями делового мира в качественных машинных переводчиках, но также и чисто научными стимулами. Стабильно работающие экспериментальные системы МП являются опытным полем для проверки различных аспектов общей теории понимания, речевого общения, преобразования информации, а также для создания новых, более эффективных моделей самого МП.

2.1 Краткая история развития МП

Впервые идея механического перевода с одного языка на другой высказывалась еще Готфридом Вильгельмом Лейбницем (1646-1716). Им же

впервые была высказана идея универсального философско-математического языка-посредника. По представлениям Лейбница, он должен был обладать полной однозначностью для устранения полисемии. В дальнейшем немало изобретателей строили механические переводчики: в 1924 году А.Вахер в Эстонии, в 1933 г. П.П.Петров-Троянский в России и Георгий Арцруни во Франции. Совсем иной интерес к проблеме проснулся лишь после появления электронно-вычислительных машин.

Задачу использования ЭВМ для перевода текстов с одних языков на другие впервые в явном виде сформулировали А.Бут и У.Уивер в 1946 г. С 1949 года в США появляется ряд научных коллективов, разрабатывающих проблематику МП, а в 1952 году проходит первая международная конференция по МП в Массачусетском технологическом институте. В 1954 г. проводится знаменитый Джорджтаунский эксперимент: перевод с русского языка на английский. Его успех обеспечил бурное развитие работ по МП в течение последующих 10 лет.

В 1955 г. проходит первый эксперимент по МП в СССР – перевод английских текстов на русский. Позже в Москве создается система англо-русского автоматического перевода (АРАП), словарь которой включал многоаспектную лингвистическую информацию (в том числе семантические толкования слов, релевантные для этих слов грамматические правила, сведения о лексической сочетаемости и др.) и насчитывал примерно по 10000 английских и русских словарных статей, а грамматика – порядка 500 правил. Система не была реализована из-за отсутствия технических возможностей, но ее идея была использована в более поздних разработках (в частности в системе ЯРАП).

Ожидалось, что для перевода научно-технических текстов могут быть быстро созданы системы промышленного масштаба. Однако первые десять лет работ привели к сильному разочарованию. В 1966 г. специально созданный Комитет национальной академии наук США ALPAC сделал официальное заключение о нерентабельности МП по сравнению с обычным переводом. Работы в области МП пошли на убыль, но не прекратились, и к середине 70-х годов сложился тот тип СМП, который лежит в основе почти всех последующих разработок.

Еще в конце 60-х годов были созданы первые системы для промышленной эксплуатации. Так, в Комиссии по атомной энергии США для получения «черновых» переводов на английский язык работ русских исследователей уже с 1964 года стала применяться система GAT, а в Отделе зарубежных технологий ВВС США – разработанная фирмой IBM система Mark II; в 70-х гг. обе системы были заменены созданной на базе GAT системой SYSTRAN.

В 70-х гг. возобновились работы по МП и в СССР. В Москве в 1974 году работы по МП начались в ИНФОРМЭЛЕКТРО, где были созданы СМП с французского на русский (ЭТАП-1) и с английского на русский (ЭТАП-2). В 1974 году открывается Всесоюзный центр переводов научно-технической литературы и документации ГКНТ и АН СССР (ВЦП), где была начата разработка трех систем промышленного масштаба: англо-русской (АМПАР), немецко-русской (НЕРПА) и французско-русской (ФРАП).

С 1975 года начинается бурное развитие СМП в Японии. К началу 90-х годов там уже работает или разрабатывается около трех десятков систем.

2.2 Периодизация и классификация систем МП

СМП можно классифицировать по нескольким основания:

1. принятый в системе **тип лингвистической стратегии** (каждый тип характеризуется преимущественным развитием систем того или иного типа):
   - *конец 40-х – середина 60-х годов:* системы первого поколения Преимущественное развитие *прямых систем МП*, реализующих «лобовое» решение проблемы перевода (близки к пословному переводу).
   - *середина 60-х – середина 70-х годов:* СМП второго поколения. Происходит интенсивное развитие синтаксических теорий и разработка СМП на их основе.
   - *середина 70-х – середина 80-х годов:* экстенсивное развитие СМП. Хорошо освоен морфологический и синтаксический анализ, но выхода к переводу через семантические структуры (СМП третьего поколения) не произошло. Развиваются интерактивные СМП, комбинирующие труд человека и ЭВМ, перевод текстов, настроенных на узкую *предметную область* (**ПО**).
   - *со второй половины 80-х годов:* резко возрастает интерес к СМП, появляется все больше многоязычных систем, используются базы знаний, привлекаются разные семантические теории.

Таким образом, в системах первого поколения операция перевода требует минимума преобразований – все элементы исходного текста, найденные в словаре, заменяются переводными эквивалентами. При этом учитывается лишь локальный контекст, он же позволяет выделять более сложные единицы перевода – обороты. Для этих систем характерны бинарность и отсутствие промежуточных структур. В системах второго поколения переводные соответствия устанавливаются не «прямым» способом, а после того, как для предложения выявлена его синтаксическая или семантическая структура.

В 80-е годы выделяется класс СМП, основанных на знаниях. В системах этого класса в качестве отдельного компонента используются *экстралингвистические знания* (знания о ПО).

Следующие два основания классификации примыкают к лингвистическому:

1. по **количеству привлекаемых языковых пар**:
   - *двуязычные* (перевод осуществляется только для данной пары языков);
   - *многоязычные*:
     - *бинарные* (анализ входного языка ведется в категориях выходного)

- *универсальные* (устройство анализа не зависит от выходного языка).

2. по **тематической ориентации**:
   - *монотематические* (настроенные на одну ПО);
   - *политематические* (охватывающие несколько ПО).

Классификация СМП может учитывать также технологические характеристики:

- **масштабность и степень реализованности** (*промышленные, развивающиеся и экспериментальные*),
- **долю участия** человека в процессе МП (*полностью автоматический перевод или человекомашиный перевод*)

2.3 Лингвистическое обеспечение систем МП

Процесс МП представляет собой *последовательность преобразований, применяемых к входному тексту и превращающих его в текст на выходном языке, который должен максимально воссоздавать смысл и, как правило, структуру исходного текста, но уже средствами выходного языка*.

В классических СМП, осуществляющих непрямой перевод по отдельным предложениям (пофразный перевод), каждое предложение проходит последовательность преобразований, состоящую из трех этапов: АНАЛИЗ – ТРАНСФЕР – СИНТЕЗ. Каждый этап в свою очередь представляет собой сложную систему промежуточных преобразований.

Цели этапов:

анализа – построить структурное описание входного предложения;

трансфера – преобразовать структуры входного предложения во внутреннюю структуру выходного предложения;

синтеза – построить правильное предложение выходного языка.

Лингвистическое обеспечение стандартной СМП включает: *словари*, *грамматики*, формализованные *промежуточные представления* единиц анализа на разных этапах преобразований.

Обычно различают следующие уровни АНАЛИЗА:

- *досинтаксический* анализ (морфологический анализ - МорфАн, анализ оборотов, неопознанных единиц текста и др.);
- *синтаксический* анализ (СинАн строит синтаксическое представление предложения, может также делиться на подуровни);
- *семантический* анализ (СемАн строит аргументно-предикатную структуру высказываний или другой вид семантического представления предложения и текста);
- *концептуальный* анализ (анализ в терминах концептуальных структур, отражающих семантику ПО, используется только в СМП, ориентированных на ограниченных ПО).

Синтез теоретически проходит те же уровни, что и анализ, но в обратном направлении.

Важным условием работы современных СМП является высокая модульность. С лингвистической точки зрения это означает, что анализ и следующие за ним процессы строятся с учетом теории лингвистических уровней. С технической же точки зрения модульность лингвистического обеспечения означает отделение структурного представления фраз и текстов от «постоянных» знаний о языке; отделение словарей от грамматик, грамматик от алгоритмов их обработки, алгоритмов от программ. Изменения внутри модуля не влияют на вид информации, поступающей на его вход и выдаваемой на его выходе, так что отдельные компоненты грамматик и словарей можно менять и дополнять, не меняя всей системы.

Словари анализа, как правило, одноязычные. Они должны содержать всю информацию, необходимую для включения данной лексической единицы в структурное представление. Часто различают словари *основ* (с морфолого-синтаксической информацией) и словари *словозначений* (с семантической и концептуальной информацией). Словари часто разделяют на словари *общеупотребительной* и *терминологической лексики*.

Грамматика и **словарь** задают лингвистическую модель, образуя основную часть лингвистических данных. Разделение грамматик и алгоритмов важно в практическом смысле тем, что позволяет менять правила грамматики, не меняя алгоритмов (и соответственно программ), работающих с грамматиками.

Основные типы грамматик:

1. *контекстно-свободные* – КС (модель языка – грамматика с конечным числом состояний, а алгоритм должен обеспечить для произвольно взятого грамматически правильного предложения дерево его вывода по правилам грамматики, и если таких выводов несколько, перечислить их);
2. *грамматики синтаксического уровня*:
   - *цепочечная грамматика* (фиксирует порядок следования элементов, то есть линейные структуры предложения, задавая их в терминах грамматических классов слов (артикль + сущ. + предлог) или в терминах функциональных элементов (подлежащее + сказуемое)). Эта реализация на ЭВМ грамматик с конечным числом состояний;
   - *грамматика (непосредственных) составляющих* – НСГ (набор правил подстановки, или исчисление продукций типа A→B… C.) Подобные грамматики порождающего типа могут использоваться как при анализе, так и при синтезе: предложения порождаются многократным применением таких правил;
   - *грамматика зависимостей* (ГЗ) задает иерархию отношений элементов предложения (главное определяет форму зависимых);
   - *категориальная грамматика Бар-Хиллела* – это версия НСГ, в ней только две категории – предложения S и имени n. Остальные определяются в терминах способности комбинироваться с этими главными в структуре непосредственных составляющих (НС).
   - *грамматики метаморфозы* - учитывающие контекстные условия (грамматика обобщенных составляющих - ГОС, расширенные сети

переходов - РСП, лексико-функциональная грамматика - ЛФГ и др. – расширения КС-правил. Условия расширяют исходные жесткие правила, придавая грамматике гибкость).

3. *унификационные грамматики* (УГ) – следующий этап обобщения анализа: способны воплощать грамматики различных видов. УГ содержит четыре компонента: пакет унификаций, интерпретатор для правил и лексических описаний, программы обработки направленных графов, анализатор с помощью граф-схемы. УГ объединяют грамматические правила со словарными статьями, синтаксические валентности с семантическими.

Центральной проблемой любой системы анализа естественного языка является проблема выбора вариантов. В НС-грамматиках применяют *фильтровый* и *эвристический* методы. Фильтровый метод состоит в том, что сначала получают все варианты анализа предложения, а затем отбраковывают те, которые не удовлетворяют некоторой системе условий-фильтров. При использовании эвристического метода с самого начала строится лишь часть вариантов, наиболее правдоподобных с точки зрения заданных критериев.

Основные идеи метода фильтров были сформулированы в работах И.Лесерфа [Лессерф 1963] и развиты рядом других исследователей, среди которых особого упоминания заслуживает Л.Н. Иорданская [Иорданская 1963], [Иорданская 1964]. Принципы работы этого метода можно суммировать следующим образом: задача заключается в том, чтобы найти синтаксическую структуру предложений, то есть сопоставить каждому содержащемуся в тексте правильной синтаксической структуры начинается с рассмотрения некоторого множества допустимых решений – связных синтаксических представлений (то есть деревьев зависимостей, пришедших с уровня синтаксического анализа), каждое из которых является гипотезой о возможных синтаксических функциях словоформ предложения. Гипотезы проверяются с помощью специальной программы фильтрации, и те из них, которые себя не оправдывают, отбрасываются. Программа служит, таким образом, в качестве своего рода фильтра, задерживающего все неправильные решения. На выход пропускается наилучшая в некотором смысле гипотеза (или несколько лучших гипотез), которая и считается решением. Возьмем в качестве примера максимальное множество гипотез, которые можно выдвинуть при анализе французской фразы *Le pilote ferme la porte «Пилот закрывает дверь»* [Лессерф 1963]. В машинной памяти, хранящей грамматику и словарь данного языка, словоформам этой фразы сопоставлены следующие информации*:*

Le	pilote	ferme	la	porte
Артикль Местоимение	Существит. Прилагательное Глагол	Существит. Прилагательное Глагол Наречие	Артикль Мест-е Существит	Существит. Прилагат. Глагол

«Для того чтобы представить эту фразу в виде «дерева зависимостей», необходимо проверить каждую из гипотез о грамматической функции каждого отдельного слова. А это подразумевает в свою очередь выбор между 2*3*4*3*3=216 грамматическими гипотезами» [Лессерф 1963], которые и образуют в совокупности множество допустимых решений. Среди допустимых решений для данной фразы есть два верных: 1. «Пилот закрывает дверь»; 2. «Сильный пилот несет ее». Чтобы отсеять неправильные решения, используются грамматические фильтры, роль которых могут выполнять, например, правила сочетаемости словоформ в составе предложения, характерные для данного языка. В частности, достаточно одного правила о невозможности последовательности «артикль + глагол в личной форме», чтобы отсеять 56 гипотез из 216.

СЕМАНТИЧЕСКИЙ УРОВЕНЬ гораздо меньше обеспечен теорией и практическими разработками. Традиционной задачей семантики считается снятие неоднозначности синтаксического анализа – структурной и лексической. Наиболее распространенный тип СемАн основан на *падежных грамматиках*, получивших популярность после появления статьи Ч.Филмора в 1968 г. В основе грамматики – понятие *глубинного*, или *семантического* падежа. Падежная рамка глагола является расширением понятия валентность: это набор смысловых отношений, которые могут (обязательно или факультативно) сопровождать глагол. Глубинные падежи (агент, адресат, цель и др.) универсальны и языковонезависимы. В пределах одного языка один и тот же глубинный падеж реализуется разными поверхностными предложно-падежными формами.

Грамматика семантического уровня в чистом виде (вне связи с синтезом) представлена в семантическом анализаторе Й.Уилкса. Его модель *семантической предпочтительности,* в которой используется тезаурусная информация, понятие псевдотекста, деление лексем на семантические примитивы, позволяет доказывать семантическую связность текста.

Другая грамматика, где не ставится задача получения полного синтаксического представления, – модель концептуальных зависимостей Р.Шенка в системе, основанной на знаниях в ограниченной ПО. Понятийные, концептуальные сети, фреймы используются как грамматика концептуального уровня.

В качестве отдельного компонента лингвистического обеспечения часто выделяются временные знания об обрабатываемом отрезке текста. В основном это сведения о внутренней структуре такого фрагмента (синтаксической, семантической и др.) – его формализованное представление, выражаемое в терминах составляющих, зависимостей, семантических сетей и др.

СИНТЕЗ в стандартных СМП трехчастного типа устроен несколько проще, чем анализ, если после работы трансфера построено одно правильное дерево фразы. Используется та же лингвистическая модель, что и в анализе, в

виде порождающей грамматики. Необходимые для построения выходной словоформы грамматические категории определяются свойствами подчиняющего узла в дереве предложения. Одноязычные словари синтеза содержат информацию, которая позволяет синтезировать требуемую словоформу; в случае, если она не может быть построена, иногда предлагаются замены.

Одна из задач синтеза – *линеаризация узлов дерева*, определяющая нужный порядок слов в синтезируемом предложении.

Рассмотрим теперь на примере нескольких отечественных СМП различные подходы к решению задач МП, в частности к созданию этапов синтеза.

Отечественные СМП ориентированы в большинстве своем на полностью автоматический перевод с последующим постредактированием. Возникающие при этом задачи решаются в разных системах разными средствами. В лингвистическом отношении здесь можно выделит три основных подхода, особенности которых наиболее четко выявляются при сопоставлении таких разработок, как комплексы АНРАП, ЭТАП и система ФРАП [Бакулов и др. 1990б].

1. Лингвистическое и программное обеспечение **комплекса АНРАП** ориентировано на максимальную технологичность: в нем используется лексикографический способ организации лингвистической информации, повышающий надежность работы комплекса. Благодаря этому он может работать с любыми типами входных текстов, в частности с грамматически неправильными предложениями и словосочетаниями, и давать какие-то – хотя бы дефектные – переводы даже при наличии во входном тексте слов, отсутствующих в используемом словаре. Обработка текстов ведется в рамках прямого подхода к переводу. Общий механизм обработки текста не привязан формальным образом к понятию предложения. Для получения перевода фразы не требуется в обязательном порядке предварительного построения ее полной синтаксической структуры: в частных случаях каждое отдельное словосочетание или даже каждое слово может переводиться отдельно. Тем самым обеспечивается устойчивость работы системы: переводной текст, хотя бы весьма несовершенный, может выдаваться даже при значительных сбоях в ее грамматическом компоненте. Это увеличивает скорость и надежность обработки текста, однако качество переводов оказывается невысоким и требуемые объемы постредактирования весьма велики.

2. **Комплекс ЭТАП** представляет собой классический вариант синтактико-семантической СМП. Он основан на широком использовании механизма межъязыковых операций и обладает развитым синтаксическим компонентом, обогащенном семантическими сведениями. Системы этого комплекса представляют собой классические СМП синтаксического поколения, состоящие из трех частей: анализ – преобразование (межъязыковые операции,

трансфер, собственно перевод) – синтез. В системе ЭТАП-1 реализован перевод через поверхностно-синтаксическую структуру. Систему ЭТАП-2, в которой перевод осуществляется на уровне нормализованных синтаксических структур, занимающих промежуточное положение между поверхностно-синтаксической и глубинно-синтаксической структурами, авторы квалифицируют как СМП 2,5 поколения [Апресян и др. 1988]. Процесс перевода делится на шесть основных этапов. Сначала идут этапы морфологического и синтаксического анализа. Затем этап нормализации синтаксической структуры. Четвертый этап – преобразование нормализованной английской структуры в нормализованную русскую. Для этого морфологические характеристики английских слов преобразуются в соответствующие русские; английские лексемы заменяются их русскими эквивалентами. Иногда это простые преобразования, а иногда сложные, взаимозависимые. Пятый этап – развертывание нормализованной структуры в синтаксическую структуру будущего русского предложения. Здесь порождается все лексико-синтаксическое своеобразие русского предложения. На долю шестого этапа – синтаксического синтеза – остаются две задачи: морфологизация синтаксической структуры и расстановка знаков препинания.

Синтаксическое представление в комплексе ЭТАП описывается деревом зависимостей, ребра которого помечены именами синтаксических отношений, а узлами являются имена лексем предложения с набором морфологических характеристик.

3. В **системе ФРАП** [Леонтьева 1986а], [Леонтьева и др. 1986б] принят подход, отличный от того, который реализован в комплексе АНРАП, и от того, на который опирается комплекс ЭТАП. Данная система проектировалась как система с семантическим языком-посредником. Синтаксический компонент в ней может не давать на выходе правильного дерева анализа. После него предусматривается этап семантической интерпретации, на вход которого могут подаваться структуры любого уровня полноты и правильности. Именно за счет работы этого этапа должно обеспечиваться приемлемое качество перевода; на этом этапе должны уточняться первоначально неправильные, неоднозначные или неполные входные структуры. В логико-алгоритмическом отношении особенность системы ФРАП, отличающая ее от комплекса ЭТАП, состоит в том, что синтаксическая информация задается и декларативными (как в указанном комплексе), и процедурными средствами.

4. Еще одна СМП, лексико-синтаксический этап которой, будет кратко рассмотрен в следующих двух абзацах, является **система PUSLAN**. В этой системе задача лексико-синтаксического (ЛС) синтеза как этапа, промежуточного между семантико-синтаксическим синтезом (при МП его роль выполняет межъязыковой перевод) и морфологическим, решается в рамках сущностного подхода к языку [Шаляпина 1999], при котором базовыми единицами описания языка являются лингвистические сущности: от конкретных (сводящиеся к элементарным или идиоматичным лексемам и аффиксам) до обобщенно-грамматических (определяющих классы частеречного

типа), а все виды правил и отношений задаются как свойства тех или иных из этих сущностей.

Представление текста, поступающее на вход системы, имеет вид цепочки ЛС-запросов, каждый из которых определяет некоторую ЛС-сущность – потенциальную русскую словоформу или функционально аналогичное ей словосочетание – как элементарную лексическую единицу, сопровождаемую ее контекстными атрибутами – морфологическими и синтаксическими. На выходе строится цепочка лексико-морфологических запросов, однозначно определяющих соответствующую последовательность синтетических русских словоформ. Их построение предусматривает не только выбор того или иного из альтернативных способов оформления русских ЛС-сущностей в зависимости от их контекста, но и коррекцию определений входных ЛС-сущностей в случае их неполноты, противоречивости и/или морфологической нереализуемости, а также осуществление лексических замен и локальных трансформаций, позволяющих вводить, устранять или заменять те или иные сущности с использованием межсловных отсылок типа лексических функций [Мельчук 1974] и их аналогов. Учитываются также такие моменты, как влияние не только синтаксических «хозяев» на свои зависимые, но и зависимых на «хозяев»; роль порядка слов как более «глубинного» выражения коммуникативной организации текста и др. [Шаляпина и др. 2001]

В следующей главе будет охарактеризована собственно система Кросслейтор, указано ее место среди других систем МП.

3. Общая характеристика и принципы работы переводчика Кросслейтор

3.1 Краткая характеристика переводчика

Задача машинного перевода требует морфологического анализа, анализа и перевода лексики (слов и словосочетаний), синтаксического анализа и синтеза, и, наконец, семантических преобразований, долженствующих обеспечить смысловое равенство введенного и выведенного предложения и/или текста в целом [Марчук 2000, с.176].

Описываемая в данной работе система МП *Кросслейтор* позволяет осуществлять перевод с каждого на каждый из следующих языков: русский, английский, испанский и турецкий. Спецификой данной системы машинного перевода является наличие в ней семантического уровня и языка-посредника, облегчающего перевод на несколько языков.

Процесс перевода разбивается на 10 основных этапов: морфологический, синтаксический и семантический анализ, фильтрацию, нахождение эквивалентов с исходного языка на язык-посредник и с языка-посредника на язык перевода, а также этапы посттрансляции, синтаксического и морфологического синтеза.

Модульность используемой методики позволяет осуществлять перевод с неограниченного количества языков. Добавление нового языка никак не затрагивает уже существующую часть переводчика, то есть для того, чтобы перевести с этого языка на любой из уже имеющихся, необходимо написать анализ данного языка, а также создать словарь «входной язык - ЯП». Для того чтобы перевести на этот новый язык, нужно лишь написать его синтез.

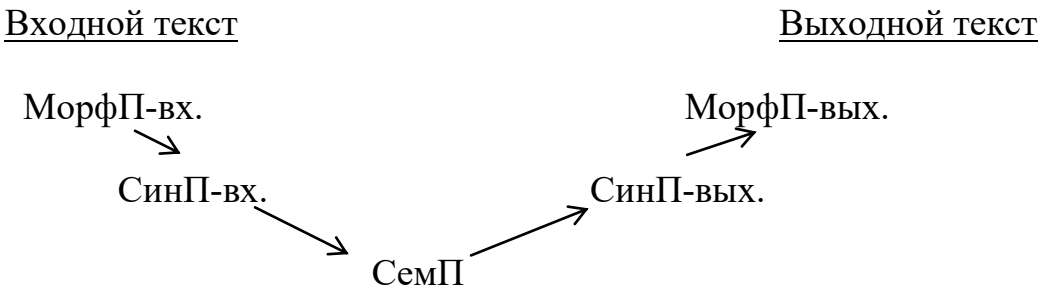
Охарактеризуем теперь компьютерный переводчик Кросслейтор в терминах приведенной выше классификации систем МП. Итак, Кросслейтор:

1. по типу лингвистической стратегии приближается к *СМП третьего поколения* (так как перевод осуществляется через синтактико-семантические структуры);
2. по количеству привлекаемых языковых пар – это *многоязычная*, *универсальная* система;
3. по тематической ориентации - *политематическая* (нет ограничений на ПО);
4. по техническим характеристикам – *развивающаяся* СМП с полностью *автоматическим* переводом.

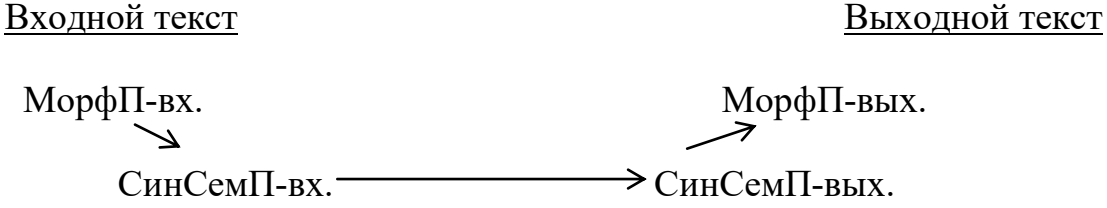
В системе машинного перевода *Кросслейтор* решаются задачи анализа входного предложения на языке оригинала и синтеза его на языке перевода. Многие современные системы перевода не использует семантику текста, а пытается перевести текст либо с использованием словосочетаний и отдельных правил грамматики языка (система СИСТРАН), либо через синтаксические структуры отдельных предложений (системы класса ЭТАП см. [Апресян и др.

1984], [Апресян и др.1989]). Однако, как показывает вся практика прикладной лингвистики, для достижения качественного или хотя бы приемлемого перевода необходим семантический анализ. Поэтому Кросслейтор объединяет в себе черты нескольких основных подходов к созданию МП. С одной стороны, в Кросслейторе, как и в комплексе ЭТАП (см. Главу 1), широко используется механизм языковых операция, развит синтаксический компонент. Перевод осуществляется на уровне нормализованных синтаксических структур, занимающих промежуточное положение между поверхностно-синтаксическими и глубинно-синтаксическими структурами. С другой стороны, в системе Кросслейтор, как и в системе ФРАП, используется язык-посредник. Отличием от системы ЭТАП является то, что в этой системе существуют два вида нормализованных структур: русские и английские, и происходит преобразование нормализованной английской в нормализованную русскую структуру. В системе Кросслейтор же существует только одна нормализованная структура, с которой и на которую осуществляются преобразования со всех четырех языков.

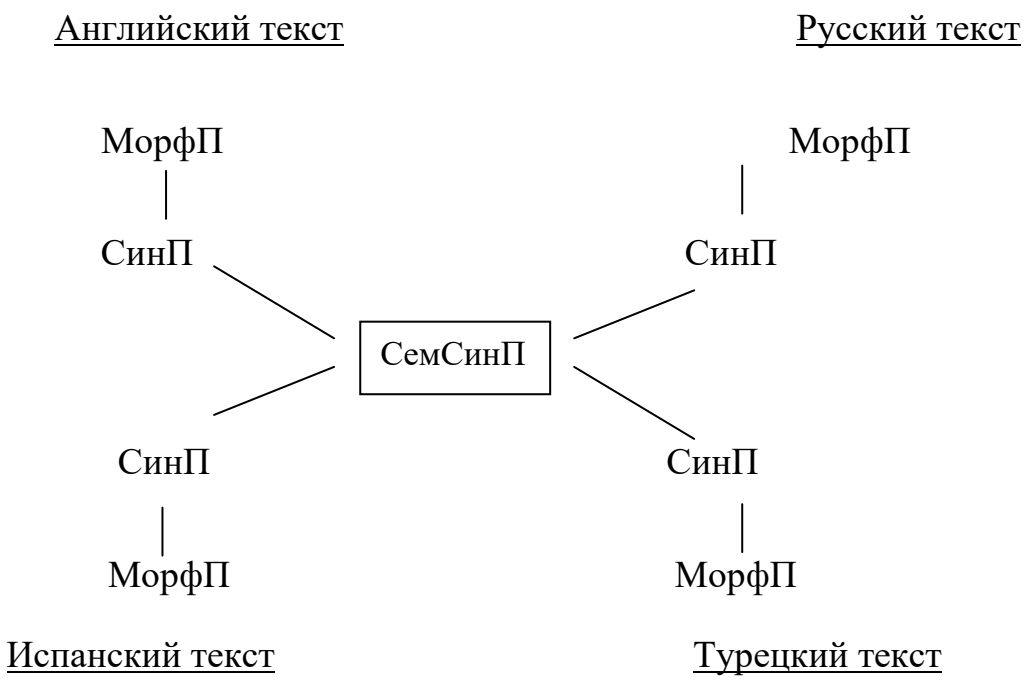
В системе Кросслейтор за основу модели СМП взята ставшая классической схема процессора автоматического перевода (АП), который предполагает последовательное построение при анализе и развертывание при синтезе трех структурных представлений – морфологического, синтаксического и семантического (МорфП, СинП, СемП соответственно) [Леонтьева 1987, с. 8]:



СемП в такой схеме предполагается единым для двух (или более) языков. Однако почти во всех реализованных системах, работающих с синтаксической структурой и учитывающих семантику, межъязыковые преобразования производятся либо через синтаксическую структуру, либо через синтактико-семантическую, то есть модель выглядит следующим образом:



Так выглядит процесс перевода, например, в системах ЭТАП. Вместо этой фактически двухкомпонентной схемы АП в системе Кросслейтор используется схема, очень близкая к классической. Для всех четырех языков она выглядит следующим образом:



На схеме не показано, но между соседними структурами (соединенными стрелками) при анализе, а иногда и при синтезе строится цепь промежуточных структур. Отличием этой схемы от классической является тот факт, что стоящее в центре схемы представление может быть названо семантическим довольно условно, так как все же имеет некоторые черты синтаксического представления.

Далее будут рассмотрены все этапы перевода по порядку.

3.2 Основные этапы работы кросслейтора

Анализ текста разбивается на следующие этапы:

1. *графематический анализ,*
2. *морфологический анализ*,
3. *словоделение сложных слов и словослияние,*
4. *синтаксический анализ*,
5. *фильтрация*,
6. *претрансляционная обработка текста (претрансляция),*
7. *постфильтрация,*
8. *этап выбора эквивалентов (с исходного языка на язык-посредник),*
9. *семантический анализ*.

Первые восемь этапов, относящиеся к этапам морфосинтаксического анализа, будут рассмотрены в этой главе, а последнему – семантическому анализу – будет посвящен отдельный раздел (см. 2.2.2). Этапы, следующие за морфологическим анализом, частично снимают омонимию, возникшую на предыдущем этапе, однако (за исключением двух этапов фильтрации, специально предназначенных для снятия омонимии) могут добавлять и новую.

Графематический анализ основан на использовании присущих данному языку законов соединения символов текста (графем). Его задача – определение границ содержательных элементов текста: словоформ, сокращений, чисел, формул, знаков препинания и т.п. Результатом графематического анализа являются сведения о границах слов внутри предложений, границах предложений и абзацев текста.

Полученные словоформы являются объектом ***морфологического анализа***. Большинство слов можно разделить на группы в зависимости от имеющихся у них парадигм. Такое деление позволяет упростить словарь, а именно: в словарной статье не вписывается каждый раз информация, касающаяся флексий, а дается только ссылка на соответствующую парадигму. Таким образом, словарь состоит из множества словарных статей, а также из общего списка парадигм. Что касается слов с нерегулярными формами словоизменения, то их парадигма приводится непосредственно в словарной статье. При анализе каждая словоформа изменяемого слова сравнивается с морфами, хранящимися в словаре, в результате чего ей сопоставляется множество основных форм слова (или одна – при отсутствии омонимии) и набор значений параметров, при котором данная нормальная форма имеет подобное написание. Например, прилагательному «красивейшая» соответствует основная форма «КРАСИВЫЙ» и следующий набор параметров: жен. род, ед. число, им. падеж, превосходная степень.

На следующем этапе ***словоделения сложных слов и словослияния*** происходит поиск словосочетаний по заданному словарю (например, сложных предлогов - «несмотря на» и др.), а также словоделение дефисных образований (например, «светло-каштановый» делится на «светлый»+«каштановый»).

Далее следует этап ***синтаксического анализа,*** который определяет связи между словами, их подчинение. На вход этого этапа поступает цепочка словоформ, каждая из которых имеет при себе все те параметры, которые уже были определены на этапе морфологического анализа (например, род, число и др.). Такое представление, поступающее на вход этапа синтаксического анализа, мы будем называть «поэлементным синтаксическим представлением». В поэлементном синтаксическом представлении отдельные словоформы и группы словоформ должны интерпретироваться «как лингвистические единицы, «понятные» модели синтаксического анализа, то есть такие, из которых может быть построено синтаксическое представление в терминах членов предложения путем их объединения в более крупные единицы с помощью синтаксических средств» [Соколова 1987]. На выходе синтаксического анализа получается «связное синтаксическое представление», представленное в виде дерева зависимостей.

Правила синтаксического анализа и синтеза в системе Кросслейтор записываются в Бэкусовских нормальных формах (в дальнейшем БНФ), фактически являющихся стандартом записи грамматик естественных, ограниченных естественных и, в большинстве случаев, специализированных

языков, таких как языки программирования, представления и хранения данных, описания форматов специальных файлов и так далее.

Для того чтобы грамматика правильно анализировала (и синтезировала) предложения входного языка, необходимо соблюдение следующих требований:

1. Принцип полноты - обеспечение возможно более полного охвата структур описываемого языка. Это возможно при условии, что тестовые предложения будут отбираться не случайно, а на основе тщательной выборки разнообразного текстового материала. Полнота может быть обеспечена, если исходные тесты хорошо представляют генеральную совокупность.
2. Принцип экономичности. Этот принцип предусматривает создание рационального количества правил во избежание ненужной детализации, повторов и избыточности описания.
3. Непротиворечивость. Необходимо четкое разграничение подобных и разных структур. Правила грамматики не должны противоречить друг другу. Логичное следование одного из другого, непересечение одних правил с другими позволит избежать ошибок.
4. Удобство и легкость записи правил. При минимальности условных символов грамматика должна нести максимум информации. Алгоритмы должны также быть достаточно гибкими для облегчения введения в грамматику новых правил для анализа/синтеза неучтенных структур.
5. Оптимизация алгоритмической обработки текста в соответствии с правилами грамматики. В эти правила должны быть заложены принципы, обеспечивающие наибольшую скорость анализа/синтеза структуры предложения..

Результатом работы синтаксического анализа одной фразы является дерево зависимостей (или набор таких деревьев в случае омонимии). В узле дерева стоит слово, при котором указаны его морфологические параметры и роль в предложении (субъект предложения, предикат и т.д.). Заданные правила построения предложения позволяют отсеять ряд морфологических омонимов. Неоднозначность (ситуация, когда одно слово может быть воспринято в нескольких значениях) является одной из основных проблем при машинном переводе. Обычно выделяют морфологическую, синтаксическую и лексическую неоднозначность. Под *морфологической* неоднозначностью понимается ситуация, когда одному и тому же написанию слова соответствует несколько уникальных наборов (начальная форма слова, параметры). Примером могут служить слова «стекло» {существительное в им.п. ед. числа «стекло», глагол пр. вр., ср рода «стекать»} и «кошки» {одушевленное существительное «кошка» в род. падеже ед. числа или неодушевленное существительное «кошки» в им./вин. падеже мн. числа}. Морфологическая неоднозначность возникает на этапе морфологического анализа и частично снимается на этапе синтаксического анализа, правила которого жестко ограничивают возможные

интерпретации. Так, например, если после этапа морфологического анализа возможны две интерпретации словосочетания «прозрачное стекло»: «ПРОЗРАЧНЫЙ (прил., ед.ч, ср.р,…) + СТЕКЛО (сущ. ср. р. ед.ч.)» и «ПРОЗРАЧНЫЙ (прил., ед.ч, ср.р,…) + СТЕКАТЬ (глаг, пр. вр., ср.р.,…), то на следующий этап перейдет только первый вариант, так как нет такого правила, в котором от глагола может зависеть прилагательное. Или другой пример: в английском языке определения, выраженные прилагательными, должны стоять *перед* определяемым словом. Если в результате разбора прилагательное, стоящее *после* глагола, не может интерпретироваться как его определение, такой вариант разбора является неверным и отбраковывается.

Синтаксическая неоднозначность (или омонимия) предполагает, что одно и то же слово на данном месте в предложении может играть разные роли, подчиняться разным словам или управлять ими различным образом. По классификации А.В. Гладкого все случаи синтаксической омонимии можно разделить на три типа: разметочная омонимия, стрелочная и конституентная омонимия [Гладкий 1985]. Для снятия синтаксической неоднозначности в системе Кросслейтор используются ***этапы фильтрации,*** базирующиеся на *эвристических* методах. Фильтрация начинается с рассмотрения некоторого множества допустимых связных синтаксических представлений (деревьев зависимостей, пришедших с уровня синтаксического анализа), каждое из которых является гипотезой о возможных синтаксических функциях словоформ предложения. Гипотезы проверяются при помощи специальных правил фильтрации, и те гипотезы, которые себя не оправдывают, отбрасываются. На выход пропускается наилучшая в некотором смысле гипотеза (или несколько, обычно не более 3-4, лучших гипотез), которая и передается на следующий этап анализа. Правила фильтрации могут просто запрещать какие-либо поддеревья (например, сочетания «любое существительное + more /более/»), либо запрещать с условием, что существует дерево с другим вариантом разбора. Например: если есть вариант разбора (связное синтаксическое представление), где английское “back” является наречием («назад»), то вариант, где “back” существительное («спина, задняя часть»), запрещается.

На ***этапе претрансляции***, который следует за этапом фильтрации, происходит *нормализация синтаксической структуры*: определяются актанты глаголов, уточняется роль слова в предложении, приводятся к стандартному виду специфические конструкции входного языка, то есть дерево зависимостей приводится к нормализованному виду, общему для всех четырех языков. Уточнение параметров слова или его роли в предложении может вестись за счет взаимосвязей слов, значений параметров других слов и прочих особенностей языка.

Следует заметить, что данный этап не уменьшает неоднозначность, а может лишь увеличить ее. Это связано с тем, что одна и та же комбинация слов может трактоваться по-разному в зависимости от контекста (семантики). Так, например, словосочетание «на + существительное» может означать

направление движения, местонахождение, срок действия и цель действия («ехать на юг на поезде на конференцию на неделю»). Подобного рода многозначность позволяет устранить семантический анализ. При использовании семантических значений слов становится ясно, что нахождение в пространстве может быть выражено парой «на + существительное места», длительности во времени – «на + существительное времени» и т.д.

Этап выбора эквивалентов разбивается на два: сначала происходит поиск и выбор эквивалентов с исходного языка на язык-посредник, а затем, уже после этапа семантического анализа, осуществляется поиск эквивалентов с языка-посредника на язык перевода (этот этап уже относится к синтезу). При этом при анализе и синтезе одного и того же языка используется одна и та же база эквивалентов. Такое разбиение, во-первых, при расширении системы (добавлении языков) делает рост количества словарей линейным, а не геометрическим. То есть, при вовлечении нового языка необходимо добавить всего один словарь (с данного языка на язык-посредник), который может использоваться в обе стороны. Во-вторых, дает возможность создать единый семантический анализ для перевода со всех языков, который будет работать с понятиями языка-посредника.

Словарная статья в каждом из четырех словарей эквивалентов, используемых для выбора эквивалентов с входного языка на ЯП и с ЯП на выходной язык, разбивается на две части – английскую (или турецкую, испанскую, русскую) и метаязыковую, в которой используются слова языка-посредника (метаязыка).

Семантический анализ

Последний этап анализа – **семантический.** Семантическое представление в Кросслейторе, так же как и в системе ФРАП [Суханова 1987], точнее назвать синтактико-семантическим. Роль этого этапа состоит в уточнении синтаксического представления: разрешении лексической неоднозначности семантическими средствами, уточнении переводных эквивалентов и др. ***Лексическая*** многозначность, т.е. когда одно и то же, с точки зрения словоизменения, слово может обозначать разные предметы, действия и так далее. Примером может служить слово «кошки» (одушевленное во мн.ч. – животное, неодушевленное – приспособление для лазания на столбы, деревья и т.п. (pluralia tantum)) или слово «коса» (прическа, инструмент и песчаная отмель). Важной особенностью этого этапа является использование ***языка-посредника (ЯП)***, уже упоминавшегося выше.

Разработанный ЯП охватывает если не весь, то большую часть литературного языка (сейчас в нем около 70 тыс. понятий). В свое время было разработано много проектов ЯП: от семоглифов до пучков грамматических соответствий, которые (как пучки, так и семоглифы) никто, включая авторов проектов, не знал, как строить и как практически применять к переводу [Мельчук и др. 1967, с. 517]. Как показали десятилетия работ по МП, трудности перехода на язык смысла от нормального естественного языка чрезвычайно

велики. Прежде всего, само понятие «смысл» так и не было точно определено. В.А. Звегинцев высказывал мысль, что предложение, взятое вне контекста, имеет не смысл, а «псевдосмысл». А разные перифразы одного и того же высказывания имеют и разные псевдосмыслы [Звегинцев 1976, с. 306]. Отсюда возникает, например, вопрос: должны ли синонимичные предложения (например: *Охотник отпустил собаку* и *Собака была отпущена охотником*) иметь одно семантическое представление или нет.

В основе ЯП, используемого в системе Кросслейтор, лежит концепция семантического пространства. Согласно этой концепции, смысл понятия интерпретируется как положение его в классификаторе (*семантическом пространстве*), имеющем шестнадцать главных рубрик (*семантических осей*). Каждая рубрика в свою очередь делится (в глубину) на более мелкие подрубрики (*семантические координаты*), уточняющие ее смысл, и так далее. Двум разным семантическим интерпретациям одной и той же лексемы соответствуют разные наборы рубрик (координат) по классификатору. Такое членение на рубрики позволяет сравнивать разложения слов (при этом происходит поиск общих осей и общих координат), что необходимо для определения степени их семантической близости.

Каждому слову ЯП соответствует набор «элементарных сем». Возьмем для примера предложение: «Программист купил новую мышь», которое содержит явную омонимию в слове «мышь». В словаре эквивалентов с русского языка на ЯП у слова «мышь» будет два значения: МЫШЬ 1 (животное) и МЫШЬ 2 (компьютерная). У этих двух слов ЯП будут разные семантические разложения:

МЫШЬ 1 определяется через такие семы, как «животное», «грызун» и др.

МЫШЬ 2 – через «компьютер», «принадлежность», «орудие».

Семантический анализатор должен выбрать лишь один вариант - наиболее правдоподобный. Для этого в довольно сложном процессе семантического анализа семантическое разложение омонимов сравнивается с семантическим разложением всех остальных слов предложения. Один из омонимов содержит элементарную сему «грызун», а другой – сему «компьютер». Слово «программист» также содержит сему «компьютер», именно это и используется для выбора между двумя омонимами слова «мышь». Конечно, когда предложение состоит из многих слов и хотя бы некоторых из них имеют по нескольку омонимов, алгоритм подобного перекрестного сличения сем «в лоб» уже не годится. Поэтому в системе Кросслейтор алгоритм работы семантического анализатора намного сложнее. Его работа основывается на понятии «семантической меры», подробно рассматриваемой в Приложении 4.

Для нас же сейчас важно лишь то, что на следующем этапе эквивалент выбирается только для слова МЫШЬ 2, что помогает избежать ошибки при переводе, которая могла бы возникнуть, если в языке перевода МЫШЬ 1 и МЫШЬ 2 обозначались бы разными словами.

Следует заметить, что этап семантического анализа является последним, который может снять омонимию. При правильно составленных правилах данного и предыдущих этапов любая пропущенная этапом семантики многозначность означает, что предложение может трактоваться и, следовательно, переводиться более чем одним предложением выходного языка.

Этапы синтеза предложений выходного языка

При синтезе предложения на языке перевода используются аналогичные этапы в обратном порядке:

1. этап выбора эквивалентов (с языка посредника на язык перевода);
2. этап посттрансляции;
3. этап синтаксического синтеза предложения;
4. образование сложных слов;
5. этап морфологического синтеза.

При этом на этапе выбора эквивалентов выходного языка, как уже было отмечено, используется та же словарная база, что и на этапе анализа, правила же посттрансляции и синтеза, хотя и имеют общие черты соответственно с этапами претрансляции и синтаксического анализа, довольно сильно от них отличаются.

Этапы посттрансляции и синтаксического синтеза будут далее подробно описаны на примере синтеза турецкого предложения, составляющего основную часть работы (см. главу 3), поэтому в данном разделе ограничимся лишь описанием их целей и очень краткой характеристикой происходящих на этих этапах преобразований.

Предыдущие этапы приводят к построению для фразы одного или нескольких деревьев зависимостей, в узлах которых стоят частично характеризованные турецкие лексемы: уже известны значения, во-первых, номинативных (например, число существительного), а во-вторых, словарных или грамматических признаков, полученные соответственно при переводе категорий и в процессе лексической подстановки.

Таким образом, для получения морфологического представления необходимо выполнить две задачи:

- приписать лексемам недостающие (а именно, чисто синтаксические, согласовательные) признаки и
- построить линейную структуру фразы выходного языка.

Первая из них решается преимущественно на этапе посттрансляции, а вторая – на этапе синтаксического синтеза. То есть, на этапе посттрансляции каждое дерево, пришедшее с этапа выбора эквивалентов, преобразуется (трансформируется) в связи с необходимостью учета специфики конструкций выходного языка.

В дальнейшем трансформированное дерево служит исходным материалом для этапа синтаксического синтеза выходного языка и будет преобразовано на этом этапе в линейную структуру. Каждый из элементов этой линейной цепочки будет уже иметь при себе все параметры, необходимые для этапа морфологического синтеза.

Предложения с неверной структурой или другими ошибками могут быть просто не разобраны на первом этапе, в результате чего система не выдаст ни одного варианта перевода. Однако и при верно построенных предложениях пока существует вероятность того, что из многих вариантов трактовки предложения система выберет неправильный вариант.

4. Анализ применимости машинного переводчика Кросслейтор для задач трансляции знаний

Для задач трансляции знаний необходимо предварительно решить большое количество лингвистических проблем. Сама по себе трансляция знаний должна осуществляться следующим образом. На первом этапе необходимо провести графематический и морфологический анализ текста с целью выделения его фрагментов и обеспечения следующего этапа – синтаксического анализа. Вместе эти этапы позволяют вычленить различные виды словосочетаний, составляющих лексику предметной области.

Словосочетания можно разбить на несколько групп. Первая – это неразрывные неизменяемые словосочетания, такие как «что-либо», «так как» и так далее. Составляющие их слова находятся рядом и не изменяются. Их можно выделить непосредственно из текста, не подвергнутого какой-либо обработке. Обычно подобные словосочетания не входят в специальную лексику. Вторая группа словосочетаний – неразрывные изменяемые. Слова в этих словосочетаниях также стоят рядом, однако могут изменяться по формам, например: «чашка Петри», «искусственный интеллект» и так далее. Для выделения таких словосочетаний необходимо предварительно провести графематический и морфологический анализы текста. И третьей группой словосочетаний являются разрывные словосочетании. Слова в таких словосочетаниях не только изменяются, но между ними могут вставляться и другие слова. Для выделения таких словосочетаний требуется провести синтаксический анализ, показывающий связи между словами.

Выделение словосочетаний является важной задачей в трансляции знаний. Сам смысл словосочетаний может быть совершенно иным, чем смысл составляющих его слов. Как это показано выше, задача выделения словосочетаний не может быть решена без применения средств анализа текстов.

Получив деревья зависимостей, отображающие связи слов, необходимо привести их к унифицированной форме, показывающей роли слов в предложении. Дело в том, что при переходе из одной предметной области в

другую обязательно производится смена терминологии. При этом оказывается, что при смене слов сменяется и их парадигма управления. То есть, например, при замене одного глагола другим необходимо сменить и предлоги, при помощи которых зависимые слова присоединяются этому глаголу. Так, например, нахождение внутри может выражаться предлогами «в» и «внутри» в зависимости от объекта («внутри биологической клетки» vs «в клетке (сооружении)»). Корректнее произвести такую смену путем выяснения роли слова на этапе анализа (с отказом от средств выражения подобной роли в данном контексте) и синтеза связи на конечных этапах.

После замены терминов входной предметной области на термины выходной необходимо провести синтез линейного текста, пригодного для чтения специалистом. Для этого требуются этапы синтаксического и морфологического синтеза.

Все описанные выше этапы уже имеются в системе машинного перевода «Кросслятор 2.0». Без каких-либо модификаций они могут войти в состав новой системы, назначением которой будет перевод терминологии из одной предметной области в другую. Для этого потребуется заменить подсистему подбора эквивалентов между различными языками на модуль подбора терминов между предметными областями.

В связи с вышеизложенным можно сделать вывод, что система машинного перевода «Кросслятор 2.0» пригодна для создания на ее основе системы трансляции знаний между различными предметными областями.

Список литературы

1. **Ахо 1978** – Ахо А., Ульман Д. Теория синтаксического анализа, перевода и компиляции Т.1, М.: Мир, 1978.
2. **Апресян 1966** – Апресян Ю.Д. Идеи и методы современной структурной лингвистики, М.: Просвещение, 1966.
3. **Апресян 1984** – Апресян Ю.Д. Лингвистическое обеспечение автоматической системы французско-русского автоматического перевода ЭТАП-1, М.: 1984.
4. **Апресян 1989** – Апресян Ю.Д. Лингвистическое описание системы ЭТАП-2, М., 1989.
5. **Бакулов и др. 1990а** – Бакулов А.Д., Леонтьева Н.Н. Теоретические основы машинного перевода//Искусственный интеллект: в 3 кн. Кн. 1 Системы общения и экспертные системы: Справочник/ под ред. Э.В.Попова, М.: Радио и связь, 1990.
6. **Бакулов и др. 1990б** – Бакулов А.Д., Леонтьева Н.Н., Шаляпина З.М. Отечественные системы машинного перевода//Искусственный интеллект: в 3 кн. Кн. 1 Системы общения и экспертные системы: Справочник/ под ред. Э.В.Попова, М.: Радио и связь, 1990.
7. **Бетин и др. 2001** – Бетин В.Н. Елкин С.В. Хачукаев Э.М. Принципы построения семантического словаря для решения задачи устранения омонимии//Вестник ВИНИТИ НТИ сер.2, 2001.
8. **Гильмуллин 2002** – Гильмуллин Р.А., Ишимов В.В. К разработке татарско-турецкого машинного переводчика//Труды международного семинара Диалог’2002 «Компьютерная лингвистика и интеллектуальные технологии», том 2, М.: Наука, 2002, (стр.133-139).
9. **Гладкий 1985** – Гладкий А.В. Синтаксические структуры естественного языка в автоматизированных системах общения. М., 1985.
10. **Звегинцев 1976** – Звегинцев В.А. Предложение и его отношение к языку и речи, М.: Изд-во МГУ, 1976.
11. **Елкин 1997** - Елкин С.В.. К вопросу об информационной физике. Часть 1.,М.: ПАИМС, 1997.
12. **Елкин и др. 2002** - Елкин С.В., Бетин В.Н., Жигарев А., Простаков О.В, Хачукаев Э.М. Разработка семантического анализа текстов при автореферировании, Вестник ВИНИТИ НТИ сер.2 .
13. **Иорданская 1963** – Иорданская Л.Н. О некоторых свойствах правильной синтаксической структуры, Вопросы языкознания, № 4, М., 1963.
14. **Иорданская 1964** – Иорданская Л.Н. Свойства правильной синтаксической структуры и алгоритм ее обнаружения, Проблемы кибернетики, вып. II, М., 1964.
15. **Клышинский и др. 2000** – Клышинский Э.С., Андреев А.С., Ёлкин С.В. Метод машинного перевода текстов// Сб. трудов 3-го научно-практического семинара "Новые информационные технологии". М.: МГИЭМ, 2000.
16. **Клышинский и др. 2002 –** Клышинский Э.С., Слезкина О.Ю. Применение модифицированных бэкусовских нормальных форм для задач анализа и

синтеза естественных языков//Новые информационные технологии: материалы пятого научно-практического семинара, М., Моск. гос. институт электроники и математики, 2002.

17. **Кононов 1953** – Кононов А.Н. Грамматика современного турецкого литературного языка, М.-Л.: Издательство Академии Наук СССР, 1953.
18. **Кулагина 1979** – Кулагина О.С. Исследования по машинному переводу, М.: Наука, 1979.
19. **Куликов и др. 1994** – Куликов В.В., Гаврилов Д.А., Ёлкин С.В. Универсальный искусственный язык Ноом-Диал, Гэлэкси Нэйшн, М., 1994.
20. **Леонтьева 1986а** – Леонтьева Н.Н. Система французско-русского автоматического перевода (ФРАП): лингвистические решения, состав, реализация// МГПИИЯ им. М.Тореза. Сборник научных трудов, вып. 271, М.,1986.
21. **Леонтьева и др. 1986б** – Леонтьева Н.Н., Кудряшова И.М., Малевич О.Б. Семантические заготовки к пониманию целого текста// МГПИИЯ им. М.Тореза. Сборник научных трудов, вып. 271, М.,1986.
22. **Леонтьева 1987** – Леонтьева Н.Н. Система французско-русского автоматического перевода (ФРАП): лингвистические решения, состав, реализация// Сборник научных трудов Выпуск 217: Машинный перевод и прикладная лингвистика. Проблемы создания системы автоматического перевода, М., 1987.
23. **Лесерф 1963** – Применение программы и модели конфликтной ситуации к автоматическому синтаксическому анализу, М.: Научно-техническая информация, 1963, вып. 10.
24. **Майзель 1957** – Майзель С.С. Изафет в турецком языке, М.-Л.: Издательство Академии наук СССР, 1957.
25. **Марчук 2000** – Марчук Ю.Н. Основы компьютерной лингвистики, М., 2000.
26. **Мельчук и др. 1967** – Мельчук И.А., Равич Р.Д. Автоматический перевода (1949-1963): Критико-библиографический справочник. М.: ВИНИТИ, 1967.
27. **Мельчук 1964** – Мельчук И.А. Опыт лингвистических моделей «Смысл ↔ Текст». М.: Наука, 1964.
28. **Мельчук 1999** - Мельчук И.А. Опыт теории лингвистических моделей «Смысл – Текст». М.: Школа «Языки русской культуры», 1999.
29. **Соколова 1987** – Соколова Е.Г. Об организации формализованного синтаксического представления в терминах членов предложения// Сборник научных трудов Выпуск 217: Машинный перевод и прикладная лингвистика. Проблемы создания системы автоматического перевода, М., 1987.
30. **Суханова 1987** – Суханова М.С. Некоторые аспекты межъязыковых преобразований и синтеза русского текста в системе ФРАП// Сборник научных трудов Выпуск 217: Машинный перевод и прикладная лингвистика. Проблемы создания системы автоматического перевода, М., 1987.
31. **Шайкевич 1995** – Шайкевич А.Я. Введение в лингвистику, М.: Изд. Российского открытого университета, 1995.

32. **Шаляпина 1999** – Шаляпина З.М. Оппозиция «часть-целое» и сущностный подход к моделированию языковой компетенции//Роман Якобсон: тесты, документы, исследования. М.: РГГУ, 1999.
33. **Шаляпина 2001** – Шаляпина З.М. Структурные валентности как универсальный инструмент описания языковой синтагматики (в рамках сущностного подхода к ее моделированию)// Московский лингвистический журнал № 5/2, М.: РГГУ, 2001 (стр. 35-84);
34. **Шаляпина и др. 2001** – Шаляпина З.М., Борисова Е.Г., Канович М.И., Панина А.С., Тарасова Е.С., Штернова О.А. Проблемы русского лексико-синтаксического синтеза при сущностном подходе к языку.