

Министерство образования
РФ

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ

Баскакова Ирина Васильевна

Тема: Исследование и разработка архитектуры системы перевода

Утверждена приказом по университету № 1276/2 от 19.05.03 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

по направлению высшего профессионального образования
552800 “Информатика и вычислительная техника”

Факультет Автоматики и Вычислительной Техники

Руководитель:

Гаврилов А.В.,

к.т.н., доц. каф. ВТ НГТУ

Новосибирск, 2003 г.

Министерство образования
РФ

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ
УНИВЕРСИТЕТ

УТВЕРЖДАЮ
Заведующий кафедрой ВТ
д.т.н., профессор
Губарев В.В.

“ ___ ” _____ 2002 г.

ЗАДАНИЕ

на магистерскую диссертацию

студенту *Баскаковой И.В.* факультета Автоматики и Вычислительной
Техники, обучающемуся по направлению 552800 “Информатика и
вычислительная техника”

Магистерская программа (специализация) 552819, “Компьютерный анализ и
интерпретация данных ”

Тема: Исследование и разработка архитектуры системы перевода

Цель работы: Исследование существующих систем машинного перевода.
Анализ технологий к построению систем машинного перевода. Разработка
структуры и основных принципов работы системы перевода с русского языка
на английский

Руководитель МД:

Гаврилов А.В.,
к.т.н., доц. каф. ВТ НГТУ

(подпись)

“ ___ ” _____ 2003 г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	5
<u>1. АНАЛИТИЧЕСКИЙ ОБЗОР ТЕХНОЛОГИЙ СОЗДАНИЯ СИСТЕМ МАШИННОГО ПЕРЕВОДА И ПОСТАНОВКА ЗАДАЧИ</u>	8
<u>1.1. ВВЕДЕНИЕ</u>	8
<u>1.2. ИСТОРИЯ РАЗВИТИЯ СИСТЕМ МП</u>	10
<u>1.3. КЛАССИФИКАЦИЯ И ТЕХНОЛОГИИ СОЗДАНИЯ СИСТЕМ МП</u>	15
<u>1.4. ВЫВОДЫ И ПОСТАНОВКА ЗАДАЧИ</u>	19
<u>2. АНАЛИЗ И РАЗРАБОТКА АРХИТЕКТУРЫ СИСТЕМЫ ПЕРЕВОДА С ПРИМЕНЕНИЕМ ТЕХНОЛОГИИ «ПАМЯТЬ ПЕРЕВОДОВ»</u>	21
<u>2.1. ОСНОВНЫЕ ЭТАПЫ ПЕРЕВОДА ТЕКСТА, С ИСПОЛЬЗОВАНИЕМ ТМ ТЕХНОЛОГИИ</u>	22
2.1.1. <i>Фрагментация и анализ текста</i>	22
2.1.2. <i>Поиск языковых пар в памяти переводов</i>	23
2.1.3. <i>Машинный перевод</i>	23
2.1.4. <i>Проверка целостности фрагментов, формата и грамматики</i>	23
<u>2.2. ПАМЯТЬ ПЕРЕВОДОВ</u>	23
2.2.1. <i>Представление данных</i>	23
2.2.3. <i>Поиск и добавление фрагментов</i>	24
2.2.4. <i>Вычисление пересечения языковых пар</i>	26
<u>2.3. ФУНКЦИОНАЛЬНАЯ СХЕМА СИСТЕМЫ МАШИННОГО ПЕРЕВОДА</u>	27
<u>Выводы</u>	27
<u>3. РАЗРАБОТКА ОСНОВНЫХ ПРИНЦИПОВ И АЛГОРИТМОВ МАШИННОГО ПЕРЕВОДА</u>	29
<u>3.1. СОВРЕМЕННОЕ СОСТОЯНИЕ СИСТЕМ МАШИННОГО ПЕРЕВОДА</u>	29
3.1.1. <i>Проект Микрокосмос</i>	29
3.1.2. <i>Системы ЭТАП-3</i>	32
3.1.3. <i>Система ФРАП</i>	38
<u>3.2. ОСНОВНЫЕ ЭТАПЫ МАШИННОГО ПЕРЕВОДА</u>	40
3.2.1. <i>Графематический анализ</i>	41
3.2.2. <i>Морфологический анализ и лемматизация</i>	42
3.2.3. <i>Фрагментационный анализ</i>	42
3.2.4. <i>Синтаксический анализ</i>	42
3.2.5. <i>Семантический анализ</i>	43
3.2.5. <i>Перевод и синтез</i>	43
<u>Выводы</u>	44
<u>4. СЕМАНТИЧЕСКИЙ АНАЛИЗ РУССКОГО ЯЗЫКА</u>	46
<u>4.1. СЕМАНТИЧЕСКИЕ СЛОВАРИ И СЛОВАРЬ РОСС</u>	47
<u>4.2. СТРУКТУРА СЕМАНТИЧЕСКОГО СЛОВАРЯ</u>	48
4.2.1. <i>Формат словарных статей</i>	48
4.2.2. <i>Семантические характеристики</i>	49
4.2.3. <i>Общая категоризация лексики (поле КАТ)</i>	51
4.2.4. <i>Семантическое отношение (поле ВАЛ)</i>	52
4.2.5. <i>Поле ДОП, НЕСОВМ</i>	53
4.2.6. <i>Поле ЛФ, ЛХi</i>	54
<u>4.3. ПОСТРОЕНИЕ СЕМАНТИЧЕСКОГО ПРЕДСТАВЛЕНИЯ ПРЕДЛОЖЕНИЯ</u>	55
4.3.1. <i>Введение</i>	55
4.3.2. <i>Построение узлов и множества словарных интерпретации узлов</i>	57
4.3.3. <i>Построения графа гипотетических связей</i>	57
4.3.3. <i>Построение и оценка древесных вариантов</i>	60
4.3.4. <i>Основные принципы работы программы семантического анализа</i>	62
<u>Выводы</u>	64
<u>5. КЛАССИФИКАЦИЯ ТЕКСТОВ</u>	65
<u>5.1. ПОСТРОЕНИЕ МОДЕЛИ СЕТИ</u>	66
5.1.1. <i>Выделение ключевых слов</i>	66
5.1.2. <i>Оценка параметров сети</i>	70
<u>5.2. ПОРОЖДЕНИЕ ТЕКСТА НА ОСНОВЕ СЕТИ</u>	72
<u>5.3. РЕШЕНИЕ И АНАЛИЗ ДЕМОСТРАЦИОННОГО ПРИМЕРА</u>	73
<u>Выводы</u>	75

	4
<u>ЗАКЛЮЧЕНИЕ</u>	76
<u>ЛИТЕРАТУРА</u>	77
<u>ПРИЛОЖЕНИЯ</u>	79
<u>ПРИЛОЖЕНИЕ 1</u>	79
<u>ПРИЛОЖЕНИЕ 2</u>	81
<u>ПРИЛОЖЕНИЕ 3</u>	82
<u>ПРИЛОЖЕНИЕ 4</u>	86
<u>ПРИЛОЖЕНИЕ 5</u>	90

Введение

Актуальность работы. Программы машинного перевода пережили недавно возрождение, во многом связанное с развитием Интернета и с ростом его доступности для всё большего числа людей.

Можно выделить два основных стимула к развитию работ по машинному переводу в современном мире. Первый – собственно научный; он определяется комплексностью и сложностью компьютерного моделирования перевода. Как вид языковой деятельности перевод затрагивает все уровни языка – от распознавания графем (и фонем при переводе устной речи) до передачи смысла высказывания и текста. Кроме того, для перевода характерна обратная связь и возможность сразу проверить теоретическую гипотезу об устройстве тех или иных языковых уровней и эффективности предлагаемых алгоритмов. Эта характеристическая черта перевода вообще и машинного перевода в частности привлекает внимание теоретиков, в результате чего продолжают возникать все новые теории автоматизации перевода и формализации языковых данных и процессов.

Второй стимул – социальный, и обусловлен он возрастающей ролью самой практики перевода в современном мире как необходимого условия обеспечения межъязыковой коммуникации, объем которой возрастает с каждым годом. Другие способы преодоления языковых барьеров на пути коммуникации – разработка или принятие единого языка, а также изучение иностранных языков – не могут сравниться с переводом по эффективности. С этой точки зрения можно утверждать, что альтернативы переводу нет, так что разработка качественных и высокопроизводительных систем машинного перевода способствует разрешению важнейших социально-коммуникативных задач.

Цель работы. Исследование существующих систем машинного перевода. Анализ технологий построения систем машинного перевода. Разработка структуры и основных принципов работы системы перевода с русского языка на английский.

Методы исследования. В качестве метода исследования был выбран сравнительный анализ программных продуктов и разработок, предлагаемых различными фирмами, исследователями в области машинного перевода и примеров реализации конкретных систем МП. Для построения сети в задаче классификации текстов использовалась теория вероятности, математическая статистика.

Научная новизна работы заключается в разработанной архитектуре системы машинного перевода, объединяющей технологии «машинный перевод» (МП) и «память

переводов» (ТМ), поскольку в настоящее время количество реальных разработок, объединяющих названные технологии, невелико.

Практическая ценность. Разработанная архитектура может стать основой для построения исследовательского варианта системы машинного перевода, который планируется использовать для анализа семантики русского языка.

Разработана программа семантического анализа предложений русского языка.

На базе этой работы планируется разработка лабораторной работы для магистерской специальности кафедры ВТ НГТУ «Интеллектуальные системы».

Апробация работы. Основные положения и результаты диссертационной работы докладывались и обсуждались на семинарах кафедры ВТ «Интеллектуальные Системы», на студенческих конференциях «Дни науки НГТУ » в 2002 и 2003 годах (г. Новосибирск), и были опубликованы в студенческом сборнике НГТУ 2003 года.

Содержание работы.

Диссертация состоит из введения, пяти разделов, заключения, списка литературы из 26 наименований и приложений. Работа содержит 76 страниц основного текста, 15 иллюстрации и 4 таблицы.

Во введении обоснована актуальность темы диссертационной работы, определены ее цель, научная новизна и практическая ценность

В первой главе приведён обзор и анализ существующих технологий построения систем машинного перевода. Рассмотрены различные классификации систем машинного перевода. Так же приведена краткая история создания систем перевода и сделан вывод о перспективах их развития.

Во второй главе изложены основные принципы перевода текста с применением ТМ технологии. Рассмотрены общие вопросы использования памяти переводов, ее преимущества и недостатки. Приведен пример построения модели памяти переводов. Предложена функциональная схема работы системы МП с использованием памяти переводов.

В третьей главе приведен аналитический обзор систем МП, а также изложены технологии машинного перевода. Рассмотрены основные этапы машинного перевода.

Четвертая глава посвящена семантическому анализу текстов на русском языке. Рассматривается алгоритм семантического анализа, основанный на использовании Русского общесемантического словаря (РОСС). Приведено описание свойств, структуры и словарных статей РОСС, используемых при семантическом анализе русского языка. Изложены этапы построения семантической структуры предложения. Описаны особенности программной реализации семантического анализа.

В пятой главе рассматривается задача тематической классификации текстов. Предложено рассматривать модель предметной области в форме сети, узлы которой представлены множеством часто встречающихся понятий текста. Описано применение законов Зипфа для построения такой сети. Показано, как оценить вероятность того, что произвольный текст был порожден на основе заданной модели сети. На основе законов Зипфа приведен пример выделения наиболее значимых для заданной темы слов.

В заключении формулируются основные выводы по результатам исследования.

1. Аналитический обзор технологий создания систем машинного перевода и постановка задачи

1.1. Введение

Перевод - это деятельность, заключающаяся в передаче содержания текста на одном языке средствами другого языка, а также результат такой деятельности. Ее теоретическим осмыслением и оптимизацией занимается дисциплина, называемая наукой о переводе и включающая в себя несколько направлений, среди которых выделяются теория перевода, анализ перевода, методика обучения переводу.

Особое место занимает машинный перевод – научная и одновременно технологическая дисциплина, связанная и с наукой о переводе, и с компьютерной лингвистикой. Машинный перевод - это выполняемое на компьютере действие по преобразованию текста на одном естественном языке в эквивалентный по содержанию текст на другом языке, а также результат такого действия.

Как и многие другие разделы прикладной лингвистики, перевод по существу междисциплинарен – он связан не только с наукой о языке, но и с литературоведением, когнитивными науками, культурной антропологией, страноведением.

Междисциплинарность теории перевода и ее практических приложений указывают на то, что перевод является не чисто языковым, а довольно сложным когнитивным феноменом. Переводя с одного языка на другой, человек использует как свои языковые знания и способности, так и самые разнообразные экстралингвистические знания (о физической природе мира, об обществе и его культуре и т.д.), причем этапы понимания и синтеза текста принципиально различаются. В самом общем виде полную схему основных языковых и когнитивных операций, сопровождающих процесс перевода с одного языка на другой (с «языка источника» L1 на «язык-цель» L2), можно представить следующим образом (см. рис. 1.1.).

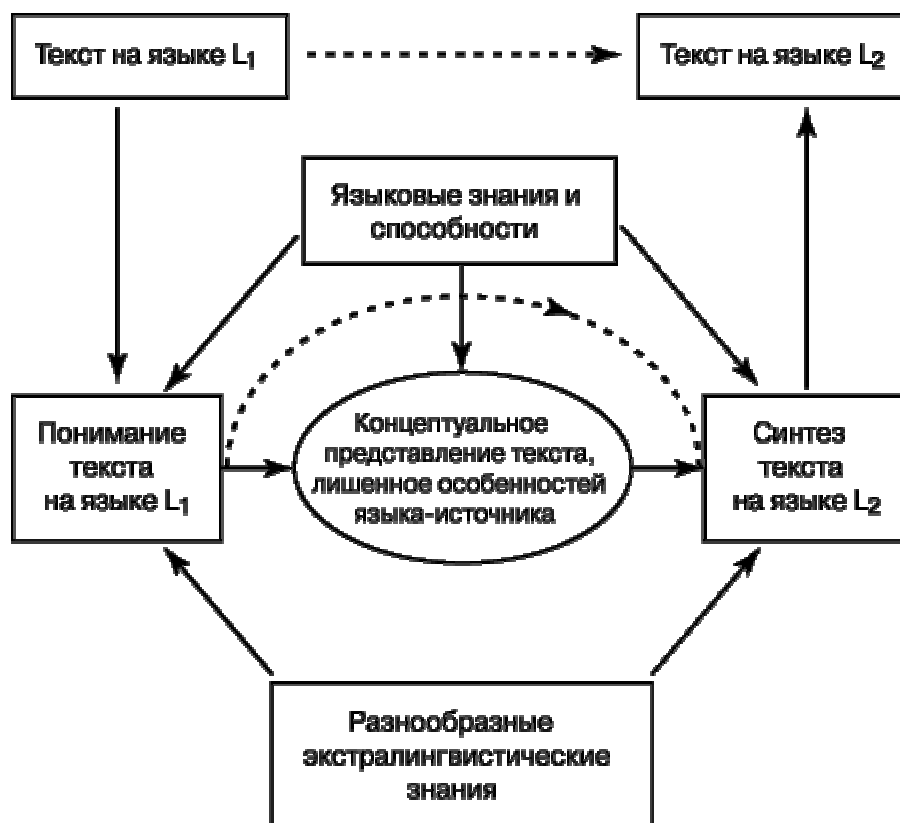


рис. 1.1. Языковые и когнитивные операции в процессе перевода

Наличие на рис.1.1 нескольких путей, ведущих от текста на L₁ к тексту на L₂, отражает тот факт, что перевод может осуществляться с разной степенью проникновения в содержание текста.

Перевод по полной схеме (сплошные жирные стрелки) предполагает, что в его ходе происходит построение некоторого концептуального (понятийного) представления содержания текста. Это представление предположительно не зависит от особенностей входного и выходного языков и учитывает всю полноту знаний, которые в связи с конкретным текстом могут быть привлечены переводчиком (или переводящим устройством) для максимально более глубокого понимания текста и максимально адекватной передачи его содержания на другом языке. Иными словами, перевод по полной схеме – это не подбор переводных соответствий, а максимально глубокое понимание текста с последующим порождением нового текста на другом языке.

Сокращенная схема перевода (жирные пунктирные стрелки) предполагает установление переводных соответствий между смыслами текстов на языке-источнике и языке-цели. При этом смыслы сохраняют особенности этих языков и никакого языка-посредника не постулируется.

Наконец, при краткой схеме перевода (жирная штрих-пунктирная стрелка) переводные соответствия устанавливаются непосредственно между выражениями языка-

источника и языка-цели.

На практике провести грань между тремя различными схемами трудно. В процессе перевода в зависимости от его задач и условий постоянно осуществляется переход от требующей наибольшей затраты когнитивных ресурсов полной схемы к наименее трудоемкой краткой через занимающую промежуточное положение сокращенную и обратно. В сфере машинного перевода относительно роли языка-посредника велись и продолжают вестись острые дискуссии [25].

1.2. История развития систем МП

Теоретической основой начального (конец 1940-х – начало 1950-х годов) периода работ по машинному переводу был взгляд на язык как кодовую систему. Пионерами МП были математики и инженеры. Описания их первых опытов, связанных с использованием только что появившихся ЭВМ для решения криптографических задач, были опубликованы в США в конце 1940-х годов. Датой рождения машинного перевода как исследовательской области обычно считают март 1947 года. Именно тогда специалист по криптографии Уоррен Уивер в своем письме Норберту Винеру впервые поставил задачу машинного перевода, сравнив ее с задачей дешифровки.

Тот же Уивер после ряда дискуссий составил в 1949 меморандум, в котором теоретически обосновал принципиальную возможность создания систем машинного перевода. Вскоре началось финансирование исследований. В 1952 состоялась первая конференция по машинному переводу, организованная логиком и математиком Й.Бар-Хиллелом.

Помимо очевидных практических нужд важную роль в становлении машинного перевода сыграло то обстоятельство, что предложенный в 1950 английским математиком А.Тьюрингом знаменитый тест на разумность («тест Тьюринга») фактически заменил вопрос о том, может ли машина мыслить, на вопрос о том, может ли машина общаться с человеком на естественном языке таким образом, что тот не в состоянии будет отличить ее от собеседника-человека. Тем самым вопросы компьютерной обработки естественно-языковых сообщений на десятилетия оказались в центре исследований по кибернетике (а впоследствии по искусственному интеллекту), а между математиками, программистами и инженерами-компьютерщиками, с одной стороны, и лингвистами – с другой установилось продуктивное сотрудничество.

В 1954 общественности были предъявлены первые результаты: фирма IBM совместно с Джорджтаунским университетом (США) успешно осуществили первый

эксперимент (вошедший в историю под названием Джорджтаунского), в ходе которого система, использовавшая словарь из 250 слов и грамматику из 6 синтаксических правил, осуществила перевод 49 заранее отобранных предложений. В том же 1954 первый эксперимент по машинному переводу был осуществлен в СССР И.К.Бельской (лингвистическая часть) и Д.Ю.Пановым (программная часть) в Институте точной механики и вычислительной техники Академии наук СССР. Первый промышленно пригодный алгоритм машинного перевода и система машинного перевода с английского языка на русский были разработаны коллективом под руководством Ю.А.Моторина. После этого работы начались во многих информационных институтах, научных и учебных организациях страны [11].

Идея машинного перевода стимулировала развитие исследований в теоретическом и прикладном языкознании во всем мире. Появились теории формальных грамматик, большое внимание стало уделяться моделированию языка и отдельных его аспектов, языковой и мыслительной деятельности, вопросам языковой формы и количественных распределений лингвистических явлений. Возникли новые направления лингвистической науки – вычислительная, математическая, инженерная, статистическая, алгоритмическая лингвистика и ряд других отраслей прикладного и теоретического языкознания. В течение 1950-х годов в учебных центрах многих стран мира были открыты отделения прикладной лингвистики и машинного перевода. Так, в СССР такие отделения были созданы в Москве, в Минском МГПИИЯ, в Ереване, Махачкале, Ленинградском университете, в университетах Киева, Харькова, Новосибирска, ряда других городов. Исследования и разработки по машинному переводу развернулись также во Франции, Англии, США, Канаде, Италии, Германии, Японии, Нидерландах, Болгарии, Венгрии и других странах, а также в международных организациях, где велик объем переводов с различных языков. В настоящее время исследования по МП ведутся и в таких странах, как Малайзия, Саудовская Аравия, Иран и др.

Исследования по машинному переводу за свою пятидесятилетнюю историю переживали как подъемы, так и спады. В начале 1960-х годов завершился первоначальный эйфорический этап в развитии МП. Этому в сильнейшей степени способствовала публикация так называемой «Черной книги машинного перевода» – доклада Специального комитета по прикладной лингвистике (ALPAC) Национальной академии наук США. В докладе была констатирована невозможность создания в обозримом будущем универсальных систем высококачественного машинного перевода. Следствием этой публикации было сокращение финансирования и общее снижение интереса к проблематике МП, однако полного сворачивания исследований, в особенности

теоретических, не произошло.

Новый подъем исследований в области МП начался в 1970-х годах и был связан с серьезными достижениями в области компьютерного моделирования интеллектуальной деятельности. Соответствующая область исследований, возникшая несколько позже МП (датой ее рождения обычно считают 1956), получила название искусственного интеллекта, а создание систем машинного перевода было осмыслено в 1970-е годы как одна из частных задач этого нового исследовательского направления.

В ходе развития идей и создания промышленных систем машинного перевода были разработаны способы автоматического морфологического анализа для основных европейских языков, методы автоматического обнаружения синтаксических структур, сформулированы требования к семантическим компонентам систем. В рамках эффективного международного сотрудничества и обмена терминологией созданы большие автоматические словари с разнообразной лексической информацией, банки терминологических данных по разным тематическим областям. Результаты работ по МП способствовали началу и развитию исследований и разработок в области автоматизации информационного поиска, логического анализа естественно-языковых текстов, экспертных систем, способов представления знаний в вычислительных системах и т.д.

В настоящее время в Российской Федерации продолжают некоторые работы по системам МП, основанным на подходе «текст-смысл-текст». Идея подхода заключается в том, что от лингвиста требуется только декларативное описание фактов языка (т.е. лингвистическая теория, претендующая, правда, на особую точность и формализованность), а алгоритмы перевода составят программист и математик. В рамках этих исследований были получены значительные теоретико-лингвистические результаты (в частности, создана теория так называемых лексических функций, нашедшая применение в лексикографии), однако для создания практических систем подобного рода подход оказался недостаточно эффективным. Все практические системы без исключения используют идею переводных соответствий, т.е. в их основе лежит модель «текст-текст» и они реализуют краткую схему перевода

Неизмеримо выросшие за последние десятилетия возможности вычислительной техники и новые программистские подходы никак не могут помочь реализовать идеи анализа и синтеза, основанные на приоритете выявления только синтаксической структуры с последующим переходом к смыслу. Выявление содержания текста в рамках человеко-машинного интерфейса может производиться, как и во всякой прикладной задаче, только с использованием как декларативных, так и процедурных знаний и при значительной опоре на лексику. Эта точка зрения обоснована, в частности, в недавних

работах отечественного специалиста по программированию и искусственному интеллекту А.С.Нариньяни.

За рубежом эксплуатируется целый ряд систем машинного перевода. Наиболее известной из их числа является система SYSTRAN, разработанная и поддерживаемая компанией SYSTRAN Software Inc. и используемая службой машинного перевода при комиссии Европейского союза. Данная служба, объем переводов в которой составляет около 2,5 млн. страниц в год, использует систему SYSTRAN для перевода с английского на немецкий, французский, испанский, греческий и итальянский языки, а также с французского на английский, испанский и итальянский. В практической эксплуатации находится ряд практических систем исследовательского центра Гренобля (Франция), систему CULT (Гонконг, ныне КНР) и ряд других. На рынке коммерческого машинного перевода предлагаются системы таких фирм, как Logos Corp., Globalinc Inc., Toshiba Corp., CompuServe и др., в том числе и санкт-петербургская компания ПроМТ, выпустившая под названием PROMT 98 усовершенствованную версию популярной системы Stylus.

Проблематика машинного перевода находит свое отражение в регулярно проводимых международных конференциях по вычислительной лингвистике COLING, а также на международных конференциях по машинному переводу MT SUMMIT.

Технические инновации 1990-х годов придали новый стимул работам по МП, привлекли в данную область новые значительные инвестиции и увенчались серьезными практическими результатами:

- появление достаточно эффективных систем машинного перевода и компьютерных словарей для работы на персональном компьютере (в том числе продуктов отечественных компаний ПроМТ, «Бит», «Арсеналь», отчасти уже упомянутых выше);
- объединение систем МП с системами оптического распознавания текста и проверка орфографии;
- создание специальных средств МП для работы в Internet, обеспечивающих либо перевод текстов на серверах соответствующих компаний, либо онлайн-перевод Web-страниц.

В сочетании с пониманием ограничений МП и реалистической формулировкой целей его использования (прежде всего, это ознакомительно-реферативные цели, что хорошо соответствует базовой идеологии Internet как средства «навигации в информационном море») все это позволяет говорить об органичном встраивании систем МП в общий процесс формирования глобального информационного общества.

Эффективность работы современной системы МП в решающей степени зависит от ее удачной настройки на конкретный подъязык (или микроподъязык) естественного

языка, на определенную лексику и ограниченный набор грамматических средств, характерных для текстов данной предметной области, а также на определенные типы документов. Учение о подъязыках с точки зрения машинного перевода было впервые сформулировано Н.Д.Андреевым (Ленинградский университет) в 1967, хотя представления о языковых регистрах, стилях, жанрах письменного текста и т.п. были хорошо известны и в традиционной лингвистике. Подъязык, с точки зрения МП, определяется в первую очередь некоторым исходным набором текстов, в рамках которого определяется входной и выходной словаря, степень распространения и характер лексической неоднозначности лексем, характер и распространенность синтаксических конструкций, способы их перевода в данной языковой паре и пр. Большую роль играют параллельные тексты и словари-конкордансы, с помощью которых можно достаточно эффективно изучить и использовать в составлении алгоритмов лексическую сочетаемость и дистрибуцию (распределение) языковых элементов в речи. Статистические характеристики подъязыков помогают упорядочить структуру соответствующих алгоритмов анализа и синтеза. Выходной словарь, ориентированный на потребности синтеза и передачи основных видов соответствий в конкретной языковой паре, обеспечивает приемлемый выходной текст. В любом из современных видов машинного перевода необходимо участие человека-редактора, удобство работы которого обеспечивается качеством и надежностью соответствующего программного обеспечения.

Перспективы развития машинного перевода связаны с дальнейшей разработкой и углублением теории и практики перевода, как машинного, так и «человеческого». Для развития теории важны результаты сопоставительного языкознания, общей теории перевода, теории закономерных соответствий, способов представления знаний, оптимизации и совершенствования лингвистических алгоритмов. Новые и более эффективные словари с необходимой словарной информацией, строгие теории терминологизации лексики, теория и практика работы с подъязыками помогут повысить качество перевода лексических единиц. Наконец, новые возможности программирования и вычислительной техники также будут вносить свой вклад в совершенствование и дальнейшее развитие теории и практики машинного перевода [4].

1.3. Классификация и технологии создания систем МП

Системы машинного перевода можно классифицировать по нескольким основаниям. Одно из них - принятый в системе тип лингвистической стратегии [18]. С этой точки зрения выделяют четыре периода, каждый из которых характеризуется преимущественным развитием систем того или иного типа.

Начальный период «бурного развития» (конец 40-х — середина 60-х годов) характеризуется преимущественным развитием *прямых систем МП*, реализующих лобовое решение проблемы перевода и обеспечивающих результаты, близкие к пословному переводу (так называемые системы *первого поколения*).

Второй период (середина 60-х— середина 70-х годов) отмечен интенсивным развитием синтаксических теорий и разработкой на их основе СМП *второго поколения*,

Третий период (середина 70-х — середина 80-х годов) можно назвать периодом экстенсивного развития СМП: они получили промышленную жизнь. Техника морфологического и синтаксического анализа хорошо освоена, на повестке дня — семантика. Но ожидаемого выхода к СМП *третьего поколения*, осуществляющего перевод через семантические структуры, не произошло. В качестве компенсации получают широкое развитие интерактивные СМП, комбинирующие труд человека и ЭВМ. Другим внешним решением семантических трудностей является ориентация на перевод ограниченных классов текстов, в частности, настроенных на узкую предметную область (ПО).

Четвертый период (со второй половины 80-х годов) характеризуется резким возрастанием интереса к МП как в практическом, так и в теоретическом плане: МП — сложная область, на которой отрабатываются новые информационные технологии. Появляется все больше многоязычных систем. Большие надежды возлагаются на мощные лексические и терминологические базы данных, базы знаний. Сближается проблематика МП и ИИ широкого профиля: в МП привлекаются семантические теории, созданные для экспертных и других систем ИИ.

Таким образом, СМП первого и второго периодов были противопоставлены прежде всего по типу лингвистической стратегии. В системах первого и полуторного поколений операция перевода требует минимума преобразований: *исходный текст* постепенно превращается в текст на *выходном, языке* путем замены всех его элементов, найденных в словаре, на переводные эквиваленты, никакая языковая модель не требуется, кроме *переводных* (в основном лексических и позиционных) *соответствий*. Учитывается лишь локальный контекст, он же позволяет собирать некоторые сложные единицы — *обороты* (поэтому такой перевод называют

пословным, или *пословно-пооборотным*). Для систем данного типа характерны *бинарность*, отсутствие промежуточных структур (анализ ведется сразу в категориях выходного языка), *одновариантность*. Их сильная сторона — простота устройства и - вытекающая отсюда большая скорость работы.

В системах второго поколения переводные соответствия устанавливаются не «прямым» способом, а только после того, как для каждого предложения выявлена в результате анализа его *синтаксическая* или синтактико-семантическая *структура* (или несколько альтернативных вариантов такой структуры). *Анализ* и *синтез* независимы (анализ, как правило, *многовариантный*, ведется в категориях входного языка, синтез — в категориях выходного), так что связь того и другого этапов обеспечивается введением особого этапа *межъязыковых операций* (собственно перевода, *трансфера*).

На этом лингвистические основания классификации СМП прервались, так как перевод через *семантический язык-посредник* (ЯП), универсальный для разных пар естественных языков, не был обеспечен единой общепризнанной лингвистической теорией. Определение и развитие СМП *третьего* и *четвертого поколений* остается делом будущего. Системы *пятого поколения* выделены в японском проекте по другому основанию: они должны стать частью компьютеров пятого поколения. Это развитые многоязычные системы, базирующиеся на самой передовой информационной технологии. Их лингвистические возможности предполагается расширить освоением речевого ввода и вывода.

В литературе 80-х годов больше говорится о делении СМП не на поколения, а на системы с *прямым* и *непрямым переводом*, а последних — на системы с *трансфером* и с *языком-посредником*. Иногда СМП классифицируют на *синтаксически-ориентированные* и *лексически-ориентированные* (или СМП, работающие *под управлением словаря* — они приближаются по технике анализа к системам класса ИИ).

В 80-е годы в отдельный класс выделяют СМП, *основанные на знаниях* (knowledge-based systems). В системах этого класса (представляющего собой подкласс систем ИИ) в качестве отдельного компонента включаются *экстралингвистические знания* (знания о ПО), хотя они могут иметь те же формы представления, что и собственно лингвистическая информация (т. е. записываться в словарях и грамматиках). Отчетливо к этому классу принадлежат те СМП, которые используют при анализе *концептуальную сеть знаний*.

Логико-алгоритмические основы современных больших наиболее развитых СМП обычно настолько гибки, что могут включать любую лингвистическую теорию и

допускают разные их комбинации. Однако они имеют, как правило, собственную модель процесса перевода.

Следующие два основания классификации примыкают к лингвистическому.

По количеству привлекаемых языковых пар СМП делятся на *двуязычные* (реализующие функцию перевода только для данной языковой пары) и *многоязычные*. Те и другие в зависимости от техники лингвистического анализа могут быть либо *бинарными* (если анализ входного языка ведется в категориях выходного), либо *универсальными* (если устройство анализа не зависит от выходного языка). Универсальная двуязычная СМП может легко стать многоязычной при комбинации с компонентами анализа и/или синтеза других универсальных систем. Система же, состоящая из совокупности бинарных СМП, может быть названа многоязычной, но не является универсальной.

По тематической ориентации различают системы *монотематические*, настроенные на одну ПО (таких большинство: TAUM-METEO, SPANAM, METAL, TITUS), и *политематические*. Иногда СМП имеют ограничения на структуру вводимых текстов (СМП с *ограниченным ЕЯ*, например, система TITUS или TITRAN, переводящая только заголовки).

Классификация СМП может учитывать также технологические характеристики: масштабность, степень реализованности, долю участия человека в процессе МП.

С точки зрения соотношения машина-человек, в 1990 году Ларри Чаилдсом была предложена следующая классификация систем машинного перевода (рис.1.2.) [26]:

- FAMT (Fully-automated machine translation) — полностью автоматизированный машинный перевод;
- HAMT (Human-assisted machine translation) — машинный перевод, сделанный при участии человека;
- MANT (Machine-assisted human translation) — перевод, осуществляемый человеком с использованием компьютера.

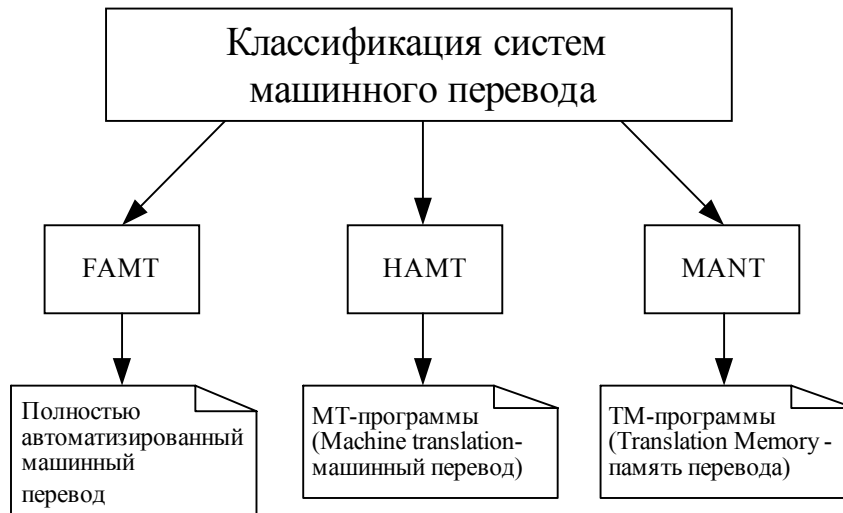


рис.1.2. Классификация систем машинного перевода

Системы типа FAMT выполняют перевод без участия человека, что не исключает ни некоторой предварительной обработки текста, ни постредактирования (как это делается и в случае обычного перевода). Но за сам процесс перевода — от ввода исходного до получения конечного текста — «ответственна» СМП, располагающая словарями, грамматикой и программами. Однако в ближайшие годы чисто машинный перевод едва ли найдет серьезное практическое применение в силу сложности, многообразия и недостаточной "формализуемости" естественных языков. Программы машинного перевода по схеме FAMT являются делом далекого будущего, поскольку в общем виде не решены проблемы автоматического понимания, перевода и синтеза текстов.

В системах типа HAMT перевод выполняет СМП, обращаясь к человеку за решением сложных случаев (снятие неоднозначности, выбор одного варианта из нескольких и т. п.).

В системах типа MANT за выполнение перевода отвечает человек, который работает за дисплеем в интерактивном режиме и может обращаться к ЭВМ, например, для поиска слов в машинных словарях и конкордансах, а также в удаленных банках терминов. Предредактирование при этом, как правило, не нужно, а постредактирование осуществляется, как при обычном переводе.

Программы категории HAMT разработчики называют MT-программы (от Machine translation - машинный перевод). Реально автоматизированный (с участием человека) машинный перевод возможен только в условиях искусственно ограниченного, как по словарному запасу, так и по грамматике, языка.

В качестве реального успешного проекта МТ-программы всегда называют немецкую систему Meteo, выполняющую перевод метеопрогнозов с французского языка на английский и обратно.

К МТ-программам относятся и продукты машинного перевода фирмы ПРОМТ, в том числе программы для просмотра содержимого Web-страниц в сети Интернет с целью поиска нужного документа.

Сравнительно недавно появился ещё один вид программ для перевода. Они основаны на технологии Translation Memory (в противоположность МТ, машинному переводу). Программы категории MANT разработчики называют ТМ-программы (от translation memory - память перевода). Эту категорию программ применяют профессиональные переводчики, осознавшие выигрыш от автоматизации их работы с помощью компьютеров.

Основу ТМ-программ составляют специализированные словари, соответствующие тематике переводимого текста. При переводе используются конструкции и значения слов и устойчивых словосочетаний, выбранные профессиональным переводчиком и занесенные в словари системы, а полученный текст подвергается интенсивному редактированию. Словари и уже переведенные фрагменты текстов, запоминаемые в ТМ-системе, могут быть повторно использованы в больших коллективных проектах, ими можно обмениваться. Поэтому ТМ-системы представляют собой важное средство автоматизации труда профессиональных переводчиков [21].

1.4. Выводы и постановка задачи

Рассмотрение существующих технологий построения систем машинного перевода позволяет сделать следующие выводы.

Полностью автоматизированный перевод, выполняемый без участия человека, все еще остается делом далекого будущего, поскольку до конца не решены проблемы автоматического понимания, перевода и синтеза текстов.

Реально автоматизированный (с участием человека) машинный перевод возможен только в условиях искусственно ограниченного, как по словарному запасу, так и по грамматике, языка. Практический машинный перевод чаще всего имеет дело с научными и техническими текстами.

Эффективность работы современной системы МП в решающей степени зависит от ее удачной настройки на конкретный подъязык (или микроподъязык) естественного языка, на определенную лексику и ограниченный набор грамматических средств,

характерных для текстов данной предметной области, а также на определенные типы документов.

В системах перевода, основанных на технологии Machine Translation (машинный перевод), перевод выполняется автоматически. Система машинного перевода обращается к человеку за решением сложных случаев (снятие неоднозначности, выбор одного варианта из нескольких и т. п.).

Системы перевода, основанные на технологии Translation Memory (память перевода), за выполнение перевода отвечает человек. Системы ТМ предназначены для того, чтобы снять всю рутину с человека, оставив ему только интеллектуальную работу над переводом.

Долгое время системы машинного перевода и памяти переводов представляли два конкурирующих направления. В настоящее время существует тенденция создания систем машинного перевода, объединяющих в себе технологии МТ и ТМ. Однако количество реальных разработок, объединяющих технологии МТ и ТМ, невелико.

Поэтому в работе ставится задача разработки структуры и основных принципов работы системы перевода, объединяющей технологии машинного перевода и памяти переводов.

2. Анализ и разработка архитектуры системы перевода с применением технологии «Память переводов»

Многие переводчики знают лишь один тип программ для повышения производительности труда: системы автоматического машинного перевода (МТ). Эти программы пытаются осуществить перевод самостоятельно без вмешательства человека. Качество автоматического перевода постоянно улучшается, однако до того времени, когда оно станет сопоставимо с человеческим, еще далеко.

Как было сказано выше, наряду с системами МТ существует другой класс программ, которые реализуют принципиально иную идею. Речь идет о технологии "Памяти Переводов" (Translation Memory, ТМ). ТМ - это программа, которая не пытается сделать работу за переводчика, а помогает ему организовать свою работу более эффективно. Большинство солидных зарубежных компаний предпочитают переводчиков, использующих эту технологию. Системы ТМ предназначены для того, чтобы снять всю рутину с человека, оставив ему только интеллектуальную работу над переводом.

Основой функционирования любой системы памяти переводов являются ранее переведенные тексты. Память переводов представляет собой базу данных, хранящую языковые пары, и определенный механизм поиска. Память переводов подразумевает просмотр ранее переведенных текстов, сравнение переводимого в текущий момент текста с тем, что хранится в базе, "вспоминает" сегменты, которые изменились незначительно, и предлагает использовать их перевод повторно. Она избавляет переводчика от необходимости по несколько раз переводить идентичные фрагменты текста, входящие в разные документы. Разумеется, критерии сходства сегментов могут быть различны, и они играют очень важную роль в расширении возможностей памяти переводов.

В то же время, использование памяти переводов требует от переводчика специальной подготовки, а также наличия соответствующего аппаратного и программного обеспечения. Другим негативным фактором является то, что для обеспечения ожидаемого эффекта все переводы должны быть сделаны в одной и той же среде, либо в средах, совместимых по формату представления данных. Наконец, полезный эффект памяти переводов проявляется с заметной отсрочкой во времени, требуя поначалу дополнительных капиталовложений.

2.1. Основные этапы перевода текста, с использованием ТМ технологии

В современных профессиональных средах перевода возможности вычислительной техники используются на различных этапах и уровнях [24]. В таблице 1 приведены основные этапы перевода текста, с использованием ТМ технологии.

Таблица 2.1

До перевода	Анализ и фрагментация текста
Во время перевода	Поиск языковых пар в памяти переводов
	Машинный перевод
После перевода	Проверка целостности фрагментов, формата и грамматики

2.1.1. Фрагментация и анализ текста

Разбиение текста на фрагменты является важным подготовительным этапом для автоматизации перевода. Задача фрагментационного анализа состоит в выделении в предложении синтаксических единств (фрагментов). Границы фрагментов не пересекают синтаксические связи, соединяющие отдельные слова или словосочетания.

Эффективность работы памяти переводов во многом определяется тем, насколько удачно решена задача фрагментации. С увеличением размера фрагментов будет уменьшаться число полных совпадений (и увеличиваться число частичных), что сильно повысит ресурсоемкость процедур поиска и потребует от переводчика значительных усилий в изучение предоставленных ему в качестве вариантов перевода языковых пар. С другой стороны, уменьшение размера фрагментов сделает их малопригодными для повторного использования, поскольку сильно возрастет влияние контекста на перевод.

Оптимальной единицей фрагментации чаще всего оказывается фрагмент предложения, ограниченный знаками препинания. Во избежание ошибочной фрагментации по точкам внутри аббревиатур и других подобных случаев используют списки исключений, словари оборотов.

Фрагментационный анализ также является одним из основных этапов машинного перевода. Он исключает возможность построения большого числа неправильных синтаксических связей, которые допускаются морфологией, синтаксисом и, возможно, семантикой (см. главу 3).

2.1.2. Поиск языковых пар в памяти переводов

Автоматическая память переводов, или просто память переводов (TranslationMemory), подразумевает, в первую очередь, просмотр ранее переведенных текстов. Она сравнивает переводимый в текущий момент текст с тем, что хранится в базе, "вспоминает" фрагменты, которые изменились незначительно, и предлагает использовать их перевод повторно. Разумеется, критерии сходства фрагментов могут быть различны, и они играют очень важную роль в расширении возможностей памяти переводов.

2.1.3. Машинный перевод

Данный способ перевода заключается в алгоритмической обработке исходного текста, в ходе которой происходит разбор фрагментов, выделяются отдельные термины и отношения между ними, после чего осуществляется замена всех терминов на соответствующие термины целевого языка в нужной форме и взаиморасположении. Технологии машинного перевода будет рассмотрена в следующей главе.

2.1.4. Проверка целостности фрагментов, формата и грамматики

Данные действия выполняются по окончании перевода и имеют своей целью проверить, все ли фрагменты остались на своих местах, сохранилась ли форматирующая информация, и корректен ли результирующий текст с точки зрения грамматики целевого языка.

2.2. Память переводов

2.2.1. Представление данных

Память переводов представляет собой базу данных, хранящую языковые пары, и определенный механизм поиска.

При построении модели памяти переводов все фрагменты разбиваются на отдельные слова (по пробелам).

В состав среды перевода помимо памяти переводов входят различные словари, связанные между собой перекрестными ссылками: общий семантический словарь РОСС, тематические словари и т.д. о которых будет сказано в 4 главе. Таким образом, систему словарей можно слить воедино с памятью переводов.

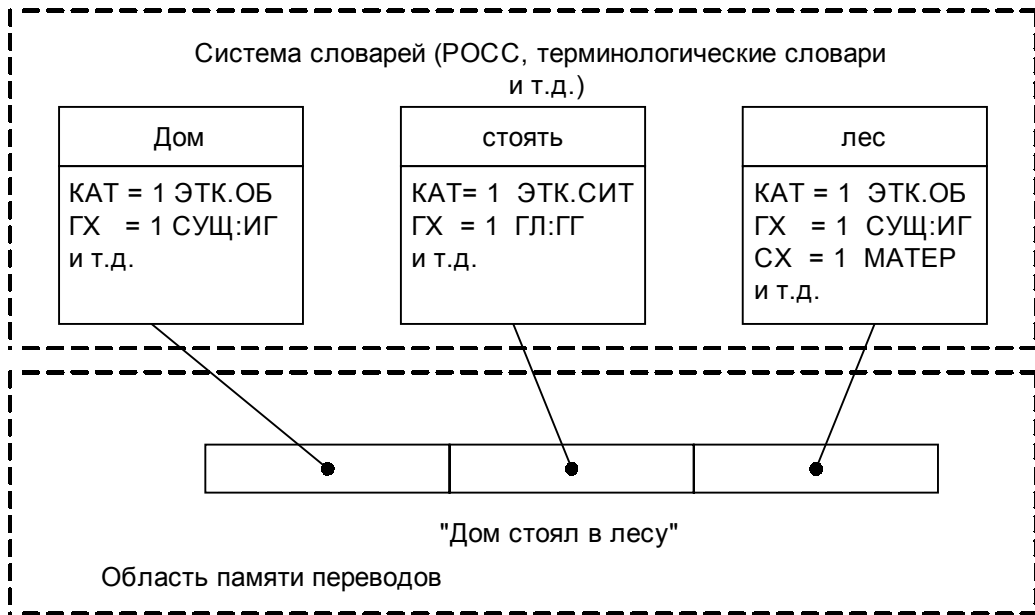


рис 2.1. Пример построения модели памяти переводов

2.2.3. Поиск и добавление фрагментов

Используя особенности выбранной структуры памяти переводов, задачу поиска фрагментов, похожих на заданный, можно решить путем выполнения следующих действий (рис. 2.2.):

- разбить заданный фрагмент на слова;
- найти в памяти переводов все узлы, соответствующие этим словам;
- спускаясь по графу отношений наследования, помещать в список найденных фрагментов все встречаемые узлы.

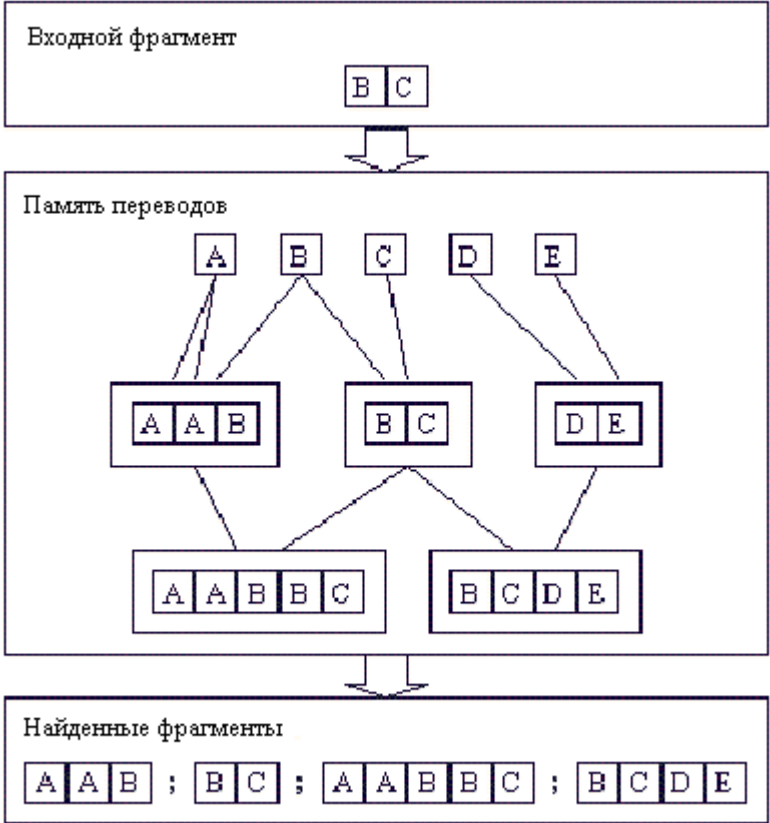


рис. 2.2. Поиск фрагментов текста в памяти переводов

Для обеспечения эффективности поиска целесообразно осуществлять оценку "пригодности" фрагментов по мере их нахождения. Например, если некоторый фрагмент полностью совпадает с эталоном, то все его потомки в графе могут быть автоматически исключены из поиска.

Условием корректности процедуры добавления является обеспечение успешного поиска. Поэтому добавляемый фрагмент должен иметь в числе своих предков (не обязательно прямых) все составляющие его слова. Среди предков должны присутствовать также узлы графа, содержащие части данного сегмента. Иными словами, если в памяти переводов присутствуют фрагменты "AB" и "CD", то фрагмент "ABCD" должен стать наследником этих двух фрагментов. Аналогично, если в памяти присутствует фрагмент "ABCD", то добавляемый фрагмент "AB" должен стать его предком. В общем случае при добавлении фрагмента в граф памяти переводов могут существовать альтернативные варианты наследования.

2.2.4. Вычисление пересечения языковых пар

Поскольку выделение общей части двух фрагментов - важный этап технологии перевода, изучим этот вопрос более детально.

Дадим определения пересечения фрагментов [7]. Пересечение фрагментов А и В - это множество фрагментов C_i , таких что:

- каждый из C_i содержится и в А, и в В;
- никакие два C_i не содержат одинаковых частей;
- не существует такого фрагмента D, что и А, и В содержат D, и D содержит один из фрагментов C_i .

Пересечение сегментов на исходном языке не обязательно изоморфно пересечению сегментов на целевом языке. Это связано с различиями правил грамматики в разных языках, порядка слов, соответствия слов понятиям.

Очень важно определить, является ли пересечение изоморфным, иными словами, можно ли считать результаты пересечения исходных и целевых фрагментов языковой парой.

Для проверки изоморфизма пересечений можно использовать подход, основанный на технологии машинного перевода. Его суть в сопоставлении терминов, образующих исходный и целевой фрагменты. Под термином понимается слово или словосочетание на заданном языке, обозначающее в этом языке конкретное понятие. Необходимо произвести грамматический разбор фрагментов с целью выделения терминов и синтаксических связей между ними. После этого можно воспользоваться словарем для определения того, какому термину в целевом сегменте соответствует заданный термин в исходном сегменте. Иными словами, изоморфизм можно определить по следующему критерию: пересечение является изоморфным, если всем терминам его исходного фрагмента, сопоставлены термины его целевого сегмента, и синтаксические связи между ними идентичны тем, которые присутствуют во фрагментах, из которых было получено пересечение.

2.3. Функциональная схема системы машинного перевода

Весь процесс работы переводчика с предлагаемой выше системой описывается схемой, изображенной на рис. 2.3. Как видно из рисунка, машинный перевод предлагается использовать в том случае, если в памяти переводов не было обнаружено фрагментов, похожих на исходный фрагмент текста.

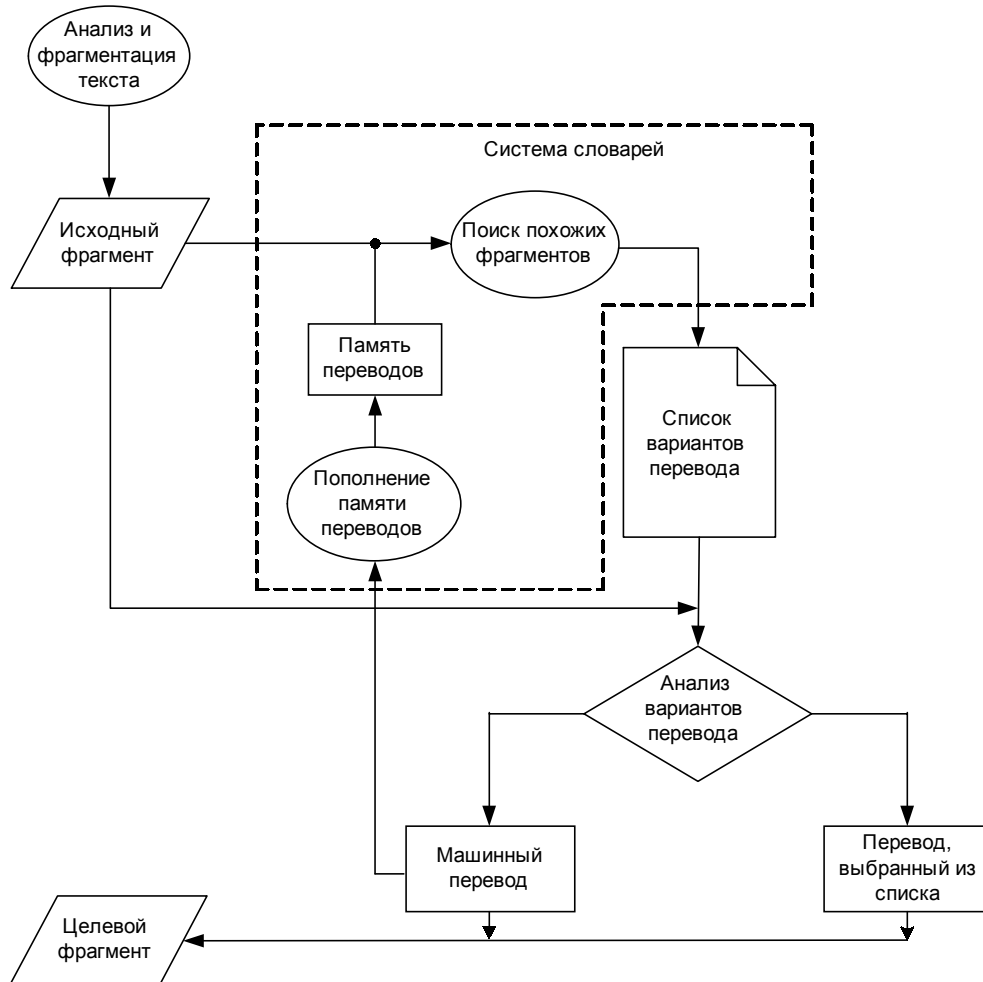


рис.2.3. Схема работы системы машинного перевода с применением технологии памяти переводов

Выводы

В данной главе выделены и рассмотрены основные этапы перевода текста, с использованием ТМ технологии:

- анализ и фрагментация текста,
- поиск языковых пар в памяти переводов,
- машинный перевод,
- проверка целостности фрагментов, формата и грамматики.

Анализ сферы применимости памяти переводов позволил выделить три условия использования рассматриваемой технологии:

- большой объем перевода,

- однотипность переводимых текстов,
- готовность к отсроченному возврату капиталовложений (полезный эффект памяти переводов проявляется с заметной отсрочкой во времени).

Поскольку в состав среды перевода помимо памяти переводов входит система словарей (РОСС, тематические словари), предложено слить воедино систему словарей с памятью переводов. Приведен пример построения такой модели

Рассмотрена задача поиска и добавления фрагментов в память переводов.

Построена функциональная схема работы системы перевода, объединяющая в себе технологии Machine Translation (машинный перевод) и Translation Memory (память перевода).

Предлагаемая модель работы системы перевода демонстрирует один из возможных вариантов построения системы, объединяющей технологии МТ и ТМ. Более того, она представляет собой попытку показать, что под машинный перевод и память переводов можно подвести общую основу, и создать такую систему профессионального перевода, в которой оба механизма действуют как единое целое.

3. Разработка основных принципов и алгоритмов машинного перевода

3.1. Современное состояние систем машинного перевода

3.1.1. Проект Микрокосмос

Проект Микрокосмос(1991-99 гг.), разрабатываемый в университете Нью-Мексико под руководством С.Ниренбурга, - одна из самых интересных и масштабных попыток использовать семантическую информацию в машинном переводе. Основные достижения разработчиков Микрокосмоса связаны именно с семантическим и послесемантическим анализами, поскольку морфологические и синтаксические анализаторы были ими заимствованы (Pangloss MT syntactic analysis module). Само название проекта Микрокосмос возникло из идеи максимально продуктивно синтезировать множество существующих на сей день теоретических разработок (т.н. микротеорий) в единую систему. К числу самых интересных микротеорий, адаптированных и улучшенных в Микрокосмосе, относят [23]:

- 1) теорию организации онтологии, принципов извлечения информации из нее;
- 2) методы применения онтологии к реальным текстам, в частности методы разрешения омонимии;
- 3) конкретные семантические микротеории.

Основным достоинством онтологического подхода в противовес бинарному переводу считается возможность более глубокого анализа текста и возможность подключать к системе перевода новые языки, не изменяя онтологии. Несмотря на это очевидное достоинство, специалисты усматривают в онтологическом подходе следующие недостатки:

- 1) Избыточность. Часто считается, что разрешить семантическую неоднозначность можно, используя несемантические методы;
- 2) Зависимость от конкретного языка. Многие полагают, что онтологии слишком сильно зависят от языковой компетенции составителя языка;
- 3) Ненаучность. Отсутствие точных методик составления онтологий делает невозможным повторение экспериментов по их воссозданию;
- 4) Инженерная сложность. Время на разработку онтологий зависит от размера онтологии не линейно, а экспонентно.

На эти контраргументы отвечает статья С.Ниренбурга [19]:

- 1) Онтология не избыточна, поскольку ни одна система машинного перевода трансферного (бинарного) типа не может полноценно справиться с задачами

восстановления кореферентных связей и метафорического переноса, без которых невозможно сделать необходимое приближение к компьютерной модели естественного языка.

2) Зависимость от конкретных языков, безусловно, - обязательное свойство онтологий, но чем больше языков описано на этой онтологии, тем более она становится независимой. Важно, что онтологии являются не чем-то принципиально ортогональным всем языкам, а неким медленно растущим образованием - общим знаменателем всех языков.

3) Контраргумент в ненаучности не лишен основания, хотя этот же довод можно отнести к любой семантической теории. По мнению Ниренбурга, точное воссоздание онтологии в каком-то другом коллективе вообще не является обязательным, поскольку их онтология имеет множество других эквивалентных альтернатив. Единственный критерий правильности онтологий - это ее практическая полезность в системах автоматической обработки текста.

4) Последний аргумент опровергается на практике. Известно, что основную сложность для перевода представляют первые 10-15% процентов лексикона (самые общие слова). Использование полуавтоматических средств для составления словарных статей дает возможность описать лексикон в 10000 слов за шесть человеко-месяцев.

Таким образом, онтологии в автоматической обработке текста становятся не слишком дорогим инструментом.

Онтология Микрокосмоса состоит из концептов и отношений между ними. Отношения записаны в слотах концепта. Концепт одновременно может содержать как абстрактную информацию (поле СЕМ), так и конкретные данные, взятые непосредственно из входного текста (поле ЗНАЧ).

Формально, концепт - это множество слотов (slots). Слот - множество пар вида <Поле, Значение> (<facet, filler>), где Поле(facet) может принимать следующие значения:

1) ЗНАЧ (Value) - значением этого поля может быть любая текстовая реализация концепта, число, буква и т.д.

2) СЕМ (Sem) - значением этого поля может быть имя другого концепта, число или шкала. Значение этого поля служит селективным ограничением для полей УМОЛЧ и ЗНАЧ. Именно с помощью поля СЕМ концепты связаны.

3) УМОЛЧ (Default) - тип значения поля такой же, как у поля ЗНАЧ. Здесь записывается значение слота по умолчанию.

4) ЕД-ИЗМ (Measuring-Unit) - здесь записывается шкала, в которой измеряется значение полей ЗНАЧ и УМОЛЧ. Шкалы являются отдельными концептами онтологии.

5) ВЕС (Salience): обозначает информационный вес концепта.

6) МАКС_ОТКЛ (Relaxable-to) - здесь записывается то, насколько селективные ограничения могут быть нарушены.

Из сказанного ясно, что значение поля может быть либо константой, отрезком шкалы, либо отсылкой к другому элементу онтологии. В первом случае поле называется атрибутом, а во втором, поскольку оно связывает два элемента, - отношением. Сами атрибуты и отношения (поля в слотах) являются концептами онтологии.

Главное отношение ВЫШЕ (is-a) тоже записывается в слотах концепта. Верхним концептом всей онтологии (по отношению ВЫШЕ) является концепт ALL. Его непосредственными потомками - концепты EVENT, OBJECT и PROPERTY. Концепт может наследоваться от многих других концептов, если они не имеют противоречащих слотов. Концепт наследует от “отцов” все слоты.

Семантическое различие между “сыновьями” может быть исключаящим или перекрывающимся.

Несколько слов нужно сказать о соотношении онтологии и лексикона в системе Микрокосмос. Лексикон содержит слова конкретного естественного языка, а онтология - концепты, которые являются общими для всех языков. Онтология и лексикон связаны отношением реализацией (instance), по которому можно сказать, какой концепт каким словом может выражаться. Отношение реализации может быть простым и с ограничениями. Ограничения могут быть у концепта (какое-нибудь значение слота равняется какому-то определенному значению) и у слов лексикона (такая-то валентность выражается таким-то грамматическим способом). В одно слово лексикона может входить много отношений реализации.

3.1.2. Системы ЭТАП-3

ЭТАП-3 – это полифункциональная система обработки текста на естественном языке, которая разрабатывается с 1980-х гг. группой российских лингвистов, математиков и программистов в Институте проблем передачи информации РАН. В основу системы ЭТАП-3 положена теория «Смысл - Текст», разработанная И.А. Мельчуком, и интегральная теория языка, разработанная Ю.Д. Апресяном.

ЭТАП-3 не является коммерческой разработкой, нацеленной на достижение конкретной прикладной цели. Его основная задача – лингвистическое моделирование естественного языка и компьютерная реализация таких моделей. Нередко в систему вводится обширная лингвистическая информация независимо от того, необходима она для повышения эффективности компьютерной обработки текста или нет.

Во всех приложениях ЭТАПА-3 используются оригинальная система трехзначной логики и детально разработанный формальный язык лингвистического описания FORET. Основной модуль ЭТАПа-3 – это система машинного перевода (МП), обслуживающая пять пар языков. Имеются системы для перевода:

- 1) с английского языка на русский,
- 2) с русского на английский,
- 3) с русского на корейский,
- 4) с русского на французский,
- 5) с русского на немецкий.

К настоящему моменту наиболее детально разработаны первые две системы. Система перевода с английского языка на русский и с русского на английский. Для остальных пар языков системы перевода существуют на уровне прототипов.

Для каждой лексемы в комбинаторном словаре приводятся ее синтаксические, словообразовательные, семантические и словообразовательные признаки, ее модель управления, а также сведения об устойчивых словосочетаниях с данной лексемой.

Если на вход ЭТАПа-3 поступает омонимичное предложение и система не может разрешить эту омонимию, то на выходе предлагаются несколько вариантов перевода. Во всех прочих случаях система выдает одну, наиболее правдоподобную, синтаксическую структуру и один, наиболее вероятный, перевод. Если же пользователь системы хочет получить все возможные переводы, он может выбрать соответствующую опцию, и система «вспомнит» все случаи неразрешенной омонимии и выдаст все возможные синтаксические структуры предложения с допустимыми для них лексическими наполнениями [1].

Язык UNL (Universal Networking Language)

Модуль UNL разрабатывается в рамках обширного международного проекта, ставящего перед собой цель преодолеть, по крайней мере, частично, языковой барьер, разделяющий пользователей Интернета.

Проект был основан в 1996 г. В настоящее время в проекте участвуют 15 университетов и научно-исследовательских институтов из Бразилии, Германии, Индии, Индонезии, Иордании, Испании, Италии, Китая, Латвии, Монголии, России, Таиланда, Франции и Японии.

Идея проекта заключается в следующем. Предлагается универсальный язык-посредник, достаточно мощный для того, чтобы на нем можно было выразить всю важнейшую информацию, которую передают тексты на естественных языках. Этот язык - Универсальный Сетевой Язык (Universal Networking Language, или UNL) предложил Х. Учида (Университет ООН). Для каждого естественного языка предлагается разработать две системы: «деконвертор», который переводил бы тексты с языка UNL на данный язык, и «энконвертор», который преобразовывал бы тексты на данном языке в выражения языка UNL. Следует подчеркнуть, что порождение текста на языке UNL не будет полностью автоматическим. Эта процедура планируется как диалог между компьютером и человеком (редактором).

В процессе интерактивного построения UNL структуры редактор будет просматривать результаты работы автоматического энконвертора, исправлять ошибки и разрешать оставшуюся многозначность. Затем редактор может запустить деконвертор и перевести отредактированное им UNL выражение на свой родной язык, чтобы проверить результаты своей работы и при необходимости внести в это выражение дополнительные изменения.

Энконвертор и деконвертор для каждого естественного языка образуют языковой сервер, который планируется разместить в Интернете. Все языковые серверы будут связаны в единую сеть UNL, что позволит пользователю Интернета переводить любой документ с UNL на свой собственный язык, а также переводить на UNL те тексты, которые он хочет сделать общедоступными.

UNL - это компьютерный язык, разработанный для представления информации в таком виде, который позволял бы порождать тексты, содержащие эту информацию, на самых разнообразных языках. Выражение языка UNL представляет собой ориентированный гиперграф, соответствующий предложению на естественном языке. Дуги графа обозначают семантические отношения, например, *agent* (деятель), *object* (объект), *time* (время), *place* (место), *instrument* (инструмент), *mode* (образ действия) и

др. В узлах графа расположены так называемые Универсальные Слова (УС) обозначающие концепты, или группы УС. Узлы могут быть снабжены атрибутами. Атрибуты содержат дополнительную информацию об использовании узла в данном предложении, например, *@imperative*, *@generic*, *@future*, *@obligation*.

Каждое УС соответствует некоторому английскому слову. Некоторые слова имеют семантические ограничители, которые уточняют значения этих слов. В большинстве случаев ограничители указывают место концепта в базе знаний. Это делается следующим образом. Универсальное Слово вида $A (icl>B)$ интерпретируется как ‘А относится к категории В’. Например, УС *coach* без каких-либо ограничителей имеет те же значения, что и английское слово *coach* в целом. Чтобы уточнить значение слова, используются ограничители. Так, выражение *coach (icl>transport)* следует понимать как ‘coach как транспортное средство’, то есть, *автобус*; выражение *coach (icl>human)* имеет интерпретацию ‘coach как человек’, то есть, *тренер*, а выражение *coach (icl>do)* – интерпретацию ‘coach как разновидность действия’, то есть глагол *тренировать*. Иными словами, аппарат ограничителей позволяет представить УС как английской слово, взятое ровно в одном значении. Кроме того, ограничители позволяют ввести концепты, для которых в английском языке отсутствуют однословные обозначения. Например, в русском языке имеется обширная группа глаголов движения, в значение которых входит указание на способ или средство перемещения: *прилететь*, *приплыть*, *приползти*, *прибежать* и др. Для глаголов этой группы отсутствуют однословные английские соответствия. Однако на основе английских слов можно построить УС, близкие им по смыслу, например, *come (met>ship)* означает ‘прибыть, причем средством передвижения является корабль’.

Приведем пример выражения на языке UNL, соответствующего английскому предложению

(3.1) *However, language differences are a barrier to the smooth flow of information in our society.*

Каждая строка UNL структуры представляет собой выражение вида *отношение (УС1, УС2)*. Для простоты семантические ограничители при универсальных словах опущены.

```

aoj(barrier.@entry.@present.@indef.@however, difference.@pl)
mod(barrier.@entry.@present.@indef.@however, flow.@def)
mod(difference.@pl, language)
aoj(smooth, flow.@def)
mod(flow.@def, information)
scn(flow.@def, society)

```

pos(society, we)

Создатели UNL планируют (при благоприятном развитии системы и достаточном финансировании) распространить сферу действия UNL на периодические издания, публикуемые в Интернете, на электронную почту и конференции, онлайн-библиотечные, научно-технические и информационно-поисковые системы, не говоря уже о публикациях таких организаций, как ООН и ЮНЕСКО [10].

Перевод с UNL на русский язык в системе ЭТАП-3

ЭТАП-3 - это трансферная система, и собственно перевод осуществляется на стадии нормализованной синтаксической структуры (НормСС). На этом уровне удобнее всего установить и соответствие между русским языком и UNL, поскольку выражения языка UNL и нормализованные синтаксические структуры обнаруживают немало общих черт. Вот наиболее существенные из них [2]:

1. Как выражения языка UNL, так и НормСС занимают промежуточное положение между поверхностным и семантическим представлениями предложения и приблизительно соответствуют так называемому глубинно-синтаксическому уровню. На этом уровне значение лексических единиц не раскладывается на примитивы, а отношения между лексическими единицами едины для всех языков;
2. Как в выражениях языка UNL, так и в НормСС узлы представляют собой терминальные элементы (лексические единицы), а не синтаксические категории;
3. Узлы содержат дополнительные характеристики (атрибуты);
4. Как в выражениях языка UNL, так и в НормСС дуги представляют собой направленные зависимости.

В то же время имеются и существенные различия между выражениями языка UNL и НормСС:

1. В НормСС все узлы представляют собой лексические единицы, а в языке UNL узел может представлять собой подграф.
2. В НормСС узел всегда соответствует одному значению слова, а значение УС может быть шире или уже, чем значение соответствующего английского слова:
 - 2.1. Значение УС может соответствовать сразу нескольким значениям одного слова (см. выше).
 - 2.2. Они могут соответствовать свободному словосочетанию (например, *computer-based* или *high-quality*).
 - 2.3. Они могут соответствовать некоторой форме слова (например, слово *best* является формой слова *good* или *well*).

- 2.4. Они могут обозначать концепт, для которого нет прямого соответствия в английском языке.
3. НормСС - это самый простой из всех связных графов, а именно, дерево, в то время как выражение языка UNL представляет собой гиперграф.
 4. В языке UNL дуги могут образовывать петли и связывать отдельные подграфы.
 5. Узлы в НормСС связаны чисто синтаксическими отношениями, не несущими никакого смысла, а отношения в языке UNL обозначают семантические роли.
 6. Атрибуты в НормСС соответствуют грамматическим характеристикам, в то время как значение многие атрибутов UNL передается лексическими средствами, как в английском языке, так и в русском (например, модальными глаголами).
 7. НормСС содержит сведения о порядке слов в предложении, а в выражении языка UNL подобная информация отсутствует.

НормСС предложения (1) выглядит следующим образом:

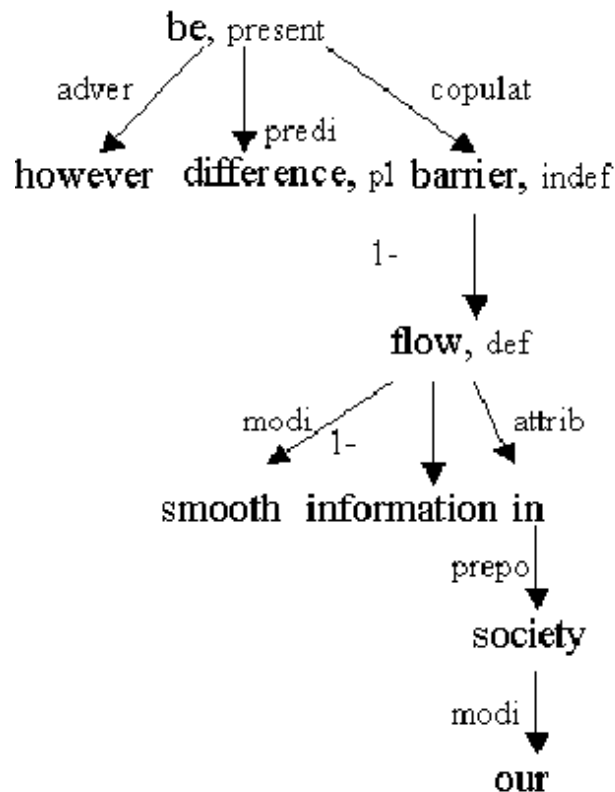


рис. 3.1. Нормализованная синтаксическая структура предложения (3.1)

Поскольку УС языка UNL обозначаются английскими лексемами, представляется целесообразным осуществить переход от представления на языке UNL к русскому предложению через посредство английской НормСС, которая будет служить промежуточным представлением (ПП). Это обеспечит наиболее простой переход от выражения на языке UNL к русскому предложению.

Таким образом, весь процесс перехода от выражения на языке UNL к русскому предложению осуществляется в три шага.

1. Переход от выражения на языке UNL к промежуточному представлению (ПП).
2. Переход от ПП к русской НормСС (НормССР).
3. Синтез русского предложения по НормССР.

Первый из этих шагов представляет собой интерфейс между языком UNL и системой ЭТАП-3, а остальные осуществляются стандартными средствами англо-русского модуля системы ЭТАП-3.

Как следует из вышесказанного, переход от выражения на языке UNL к НормСС должен решать следующие пять задач:

1. Заменить все УС английскими словами везде, где это возможно. Русские лексемы появятся на этапе англо-русского перевода при обращении к английскому словарю. Если для УС не нашлось английского эквивалента, следует выразить значение этого УС другими средствами.
2. Перевести синтаксические отношения языка UNL в синтаксические отношения ЭТАПа-3, либо непосредственно, либо с помощью лексических средств.
3. Перевести атрибуты языка UNL в грамматические характеристики ЭТАПа-3, либо непосредственно, либо с помощью лексических средств.
4. Преобразовать граф UNL в дерево зависимостей.
5. Определить порядок слов в предложении.

Первая и (отчасти) вторая задача решаются при помощи словарей UNL - английского и английского комбинаторного. За все остальные задачи отвечают правила, написанные на формально-логическом языке FORET.

Таким образом, все эти задачи решаются либо при помощи словарей, либо при помощи правил. Правила подразделяются на три класса в зависимости от степени универсальности: различаются ОБЩИЕ, ТРАФАРЕТНЫЕ и СЛОВАРНЫЕ правила. Общие правила могут активироваться при обработке любого предложения. Два других типа правил применяются только в том случае, если в обрабатываемом предложении имеется слово, которое содержит отсылку к некоторому правилу (в случае трафаретного правила) или само правило (в случае словарного правила). Подобная организация правил обеспечивает автоматическую настройку системы: активируются только те правила, которые требуются для обработки того или иного предложения.

3.1.3. Система ФРАП

Система ФРАП (французско-русского автоматического перевода) была разработана коллективом лаборатории машинного перевода Всесоюзного центра переводов совместно с коллективом лаборатории машинного перевода МГПИИЯ им М. Горького [14].

Система ФРАП была реализована в двух версиях: ФРАП1 (1976-1980) и ФРАП2(1980-1986). В системе ФРАП1 синтаксический и семантический компоненты работали в автономном режиме, между собой они не были состыкованы. В системе ФРАП2 был полностью реализован синтаксический компонент и начат семантический компонент. Устойчивый перевод осуществлялся на синтаксическом уровне, поскольку семантический анализатор не был закончен, и семантический словарь содержал всего 1000 входов. Однако именно семантическая часть ФРАП2 представляет собой интересную попытку использовать смысловые механизмы для машинного перевода.

Компоненты, составляющие переводческую модель, - лингвистические процессоры, которые друг за другом обрабатывают входной текст. Вход одного процессора является выходом другого. Выделяются следующие компоненты:

- Графематический анализ. Выделение слов, цифровых комплексов, формул и т.д.
- Морфологический анализ. Построение морфологической интерпретации слов входного текста.
- Синтаксический анализ. Построение дерева зависимостей всего предложения.
- Семантический анализ. Построение семантического графа текста.
- Информационный анализ. Соотнесение в семантического графа с внешними базами данных.

Для каждого уровня разрабатывался свой язык представления. Язык представления, как полагается, состоял из констант и правила их комбинирования. На графематическом уровне константами были графематические дескрипторы (ЛЕ – лексема, ЦК – цифровой комплекс и т.д.) На морфологическом уровне – граммы (**рд** – родительный падеж, **мн** -множественное число). На синтаксическом – названия отношений (**subj** – отношение между подлежащим и сказуемым, **circ** - обстоятельство). О других уровнях будет сказано ниже.

С каждого уровня представления можно сделать переход к такому же представлению на другом естественном языке, что позволяет осуществлять перевод, даже если “глубокие” (семантический и информационный) анализаторы не смогли обработать текст. Основой для построения уровней служили результаты работы предыдущих этапов, но, что важно, последующие анализаторы также могли улучшить представление

предыдущих. Например, для какого-то предложения синтаксический анализатор не смог построить полного дерева зависимостей, тогда, возможно, семантический анализатор сможет спроектировать им построенный семантический граф на синтаксис.

Такой многоуровневый подход позволяет предложить критерии оценки систем машинного перевода. Разработчики ФРАП (12) показали, что для достижения адекватности перевода (равенство по смыслу входному тексту) и грамматической правильности выходной фразы необходимо присутствие всех пяти этапов, причем адекватность перевода можно гарантировать только после работы «глубоких» анализаторов. Таким образом, критерии оценки систем машинного перевода сводятся к оценке проработанности отдельных уровней представления.

Перейдем к описанию семантического представления системы ФРАП.

Семантический аппарат системы ФРАП был потом использован в системах ПОЛИТЕКСТ [15], например, в Русском общесемантическом словаре (РОСС) [16].

В центре семантического аппарата ФРАП (первая часть семантики) находятся два перечня: семантических характеристик (СХ) и смысловых отношений (СО). Используется минимальное количество семантических характеристик: ВЕЩВО(«вещество»), ИЗМ(«изменение»), ИНТЕЛ(«интеллектуальность»), ИНФ(«информация») и т.д.; слова характеризуются по признаку принадлежности к одному или нескольким классам. СХ обеспечивают проверку семантического согласования при интерпретации связей в тексте. Никаких жестких критериев составления перечня СХ не существует. Перечень не однороден: некоторые СХ можно условно назвать признаками, а другие объектами. Например, ВОСПР («слышать», «видеть»), ИНТЕЛ («изучать», «решать»), ХОР («взаимопомощь», «мужество»), ГОС («республика», «министерство») и т.д. – заведомо СХ-признаки, а НОСИНФ («книга», «газета»), УСТР («компьютер», «автомобиль»), ДОЛЖ («повар», «парработник») - СХ-объекты. Характеризовать слова можно целыми формулами, составленными из СХ, например,

СХ(«компьютер»)=ИНТЕЛ,УСТР; СХ(«министр») = ГОС, ДОЛЖ.

Но даже комбинирования признаков, на самом деле, не хватает: перечень СХ-объектов заведомо неполон, т.е., его не хватает для описания всего языка.

Второй класс СХ-признаков представляет область «чистой» семантики, поскольку эти СХ фактически не используются как селективные ограничения, а скорее являются частью смысла слова. Общие СХ-признаки (АБСТР (*модель, план, структура, тенденция*), МЕСТОИМ (*проблема, вопрос, намерение*), ЭМОЦ (*мизерный, могучий, несчастный*), СОБИР (*библиотека, молодежь, группа*)) вызывают наибольшую трудность у

составителей словаря, что объясняет наибольший процент ошибок при приписывании этих СХ.

В том или ином виде основное ядро отношений системы ФРАП (АГЕНТ, ИДЕНТ (идентификатор), ПРИНАДЛ(принадлежность), АДР (адресат), СУБ (субъект, ОБ (объект), ПАЦИЕН (пациент), СОДЕРЖ (содержание), МОДЛ (модальность), КОН-Т (конечная точка), ИСХ-Т (исходная точка), СРЕДСТВО, ОГРН(ограничение), КОЛИЧ (количество) и т.д.) уже давно является частью всеобщего лингвистического аппарата. С похожими отношениями работают многие исследователи [1, 16]. Эти отношения, входящие в основное ядро, наиболее часты в словаре РОСС.

3.2. Основные этапы машинного перевода

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.

История машинного перевода насчитывает немногим более 50 лет. За это время сменилось несколько поколений систем машинного перевода - от первых программ, использовавших ограниченные ресурсы универсальных компьютеров первого поколения до современных коммерческих продуктов, использующих мощные ресурсы серверов и персональных компьютеров, включая ПК, в которых можно размещать карманные словари, а также компьютерные сети.

Современный машинный, или автоматический перевод, осуществляется с помощью человека:

- пред-редактора, который тем или иным образом предварительно обрабатывает подлежащий переводу текст,
- интер-редактора, который участвует в процессе перевода,
- пост-редактора, который исправляет ошибки и недочеты в переведенном машиной тексте.

Системы машинного перевода обычно строятся модульно, где каждый модуль принимает на вход некоторое представление текста и вырабатывает свое выходное представление.

Некоторые процессоры имеют уже устоявшиеся названия и функции, например, морфологический процессор отвечает за лемматизацию входного текста.

Основываясь на анализе вышеописанных систем МП, выделим следующие основные этапы перевода:

1. Графематический анализ. Выделение слов, цифровых комплексов, формул и т.д.

2. Морфологический анализ. Построение морфологической интерпретации слов входного текста.
3. Фрагментационный анализ. Выделение в предложении синтаксических единств (фрагментов).
4. Синтаксический анализ. Построение дерева зависимостей всего предложения.
5. Семантический анализ. Построение семантического графа текста.
6. Перевод входных словоформ (трансфер) и синтез выходных словоформ и предложения в целом на выходном языке.

Основные этапы перевода разрабатываемой системы МП представлены на рис.

Программно был реализован один из этапов машинного перевода – семантический анализ, речь о котором, более подробно, пойдет в следующей главе.

3.2.1. Графематический анализ

Графематический анализ - это начальный анализ естественного языка, представленного в виде цепочки текстовых знаков, вырабатывающий информацию, необходимую для дальнейшей обработки Морфологическим и Синтаксическим процессорами. Графематический анализ работает с внешним представлением текста.

Единицей графематического анализа является цепочка символов, выделенная с двух сторон пробелами

В задачу графематического анализа входят: разделение на слова, цифровые комплексы; выделение дат, электронных адресов URL, неизменяемых оборотов; выделение ФИО (фамилия, имя, отчество), когда имя и отчество написаны инициалами; деление на предложения, абзацы.

3.2.2. Морфологический анализ и лемматизация

Морфологический компонент осуществляет морфоанализ и лемматизацию русских словоформ.

Лемма – это нормальная форма слова. Например, для существительных – это единственное число (если оно есть у существительного), именительный падеж. То есть лемматизация - приведение текстовых форм слова к словарным; а морфоанализ – приписывание словоформам морфологической информации.

Данную задачу не представляет труда выполнить для русского языка благодаря его развитой морфологии практически со стопроцентной точностью [20].

3.2.3. Фрагментационный анализ

Задача фрагментационного анализа состоит в выделении в предложении синтаксических единств (фрагментов). Фрагменты – это главные и придаточные предложения в составе сложного, причастные, деепричастные и другие обособленные обороты.

Первая важная особенность фрагментов заключается в том, что их границы не пересекают синтаксические связи, соединяющие отдельные слова или словосочетания. Таким образом, при успешной работе фрагментационного анализа перед синтаксическим исключается возможность построения большого числа неправильных синтаксических связей, которые допускаются морфологией, синтаксисом и, возможно, семантикой.

Второе важное свойство – членение предложения на фрагменты в большинстве случаев соответствует его делению на крупные семантические узлы. Так, результат фрагментационного анализа несет и некоторую семантическую информацию о предложении.

3.2.4. Синтаксический анализ

Цель синтаксического анализа - автоматическое построение функционального дерева фразы, т.е. нахождение взаимозависимостей между разноуровневыми элементами предложения. На вход синтаксису подаются результаты морфологического анализа (каждой словоформе сопоставлено максимально возможное для данной словоформы множество морфологических интерпретаций) и фрагментационного анализа.

3.2.5. Семантический анализ

В отличие от синтаксического анализа семантический этап использует формальное представление смысла составляющих входной текст слов и конструкций. В сферу семантического анализа входит [22]:

- Построение семантической интерпретации слов и конструкций;
- Установление «содержательных» семантических отношений между элементами текста, которые уже принципиально не ограничены размером одного слова (могут быть больше или меньше одного слова).

Можно сказать, что создание полных систем машинного перевода для русского языка, использующих семантический анализ, является чрезвычайно актуальной задачей. Поэтому, в следующей главе подробнее остановимся именно на семантическом анализе русскоязычных текстов на естественном языке.

3.2.5. Перевод и синтез

На этапе трансфера решаются следующие задачи [23]:

- Для узлов русского семантического графа ищутся английские эквиваленты по английскому словарю, из которых строятся английские семантические узлы;
- Английские семантические отношения строятся по русским отношениям с необходимыми перестройками;
- Актантам английских узлов приписываются грамматические характеристики в соответствие с их словарными статьями.

На этапе синтеза, работающим непосредственно после этапа трансфера, осуществляется следующее:

- Порождаются английские словоформы по заданным на этапе трансфера граммемам;
- Определяется порядок слов;
- Осуществляется перевод терминов, групп времени, слов, не вошедших в семантические словари;
- Синтезируются артикли для именных групп.

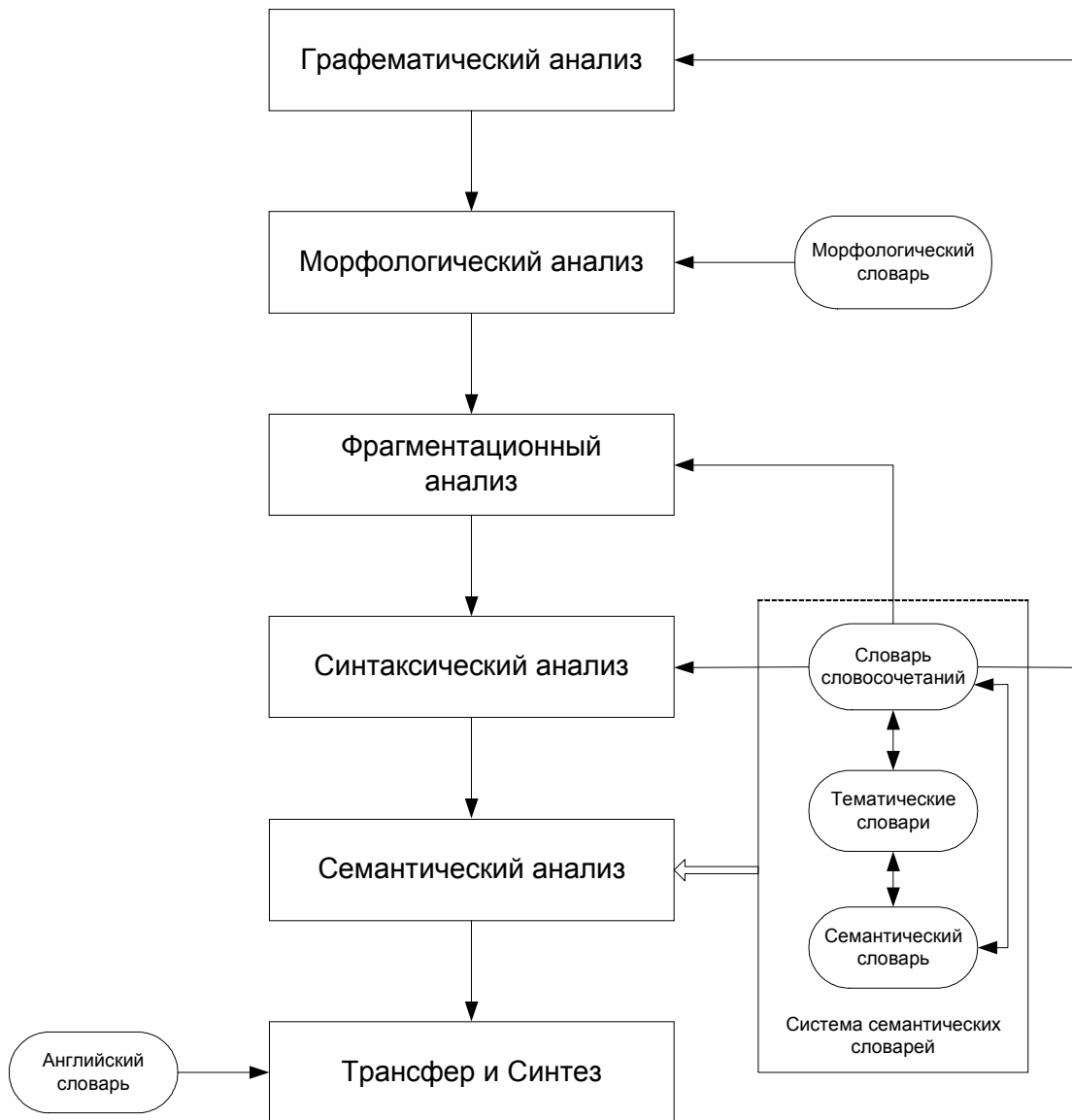


рис.3.2. Основные этапы машинного перевода

Выводы

Сделан аналитический обзор систем МП и технологий машинного перевода.

Системы машинного перевода обычно строятся модульно, где каждый модуль принимает на вход некоторое представление текста и вырабатывает свое выходное представление. Компоненты, составляющие переводческую модель, - лингвистические процессоры, которые друг за другом обрабатывают входной текст.

Основываясь на анализе различных систем МП, были выделены и рассмотрены следующие основные этапы перевода:

- Графематический анализ. Выделение слов, цифровых комплексов, формул и т.д.
- Морфологический анализ. Построение морфологической интерпретации слов входного текста.

- Фрагментационный анализ. Выделение в предложении синтаксических единств (фрагментов).
- Синтаксический анализ. Построение дерева зависимостей всего предложения.
- Семантический анализ. Построение семантического графа текста.
- Перевод входных словоформ (трансфер) и синтез выходных словоформ и предложения в целом на выходном языке.

4. Семантический анализ русского языка

Обычно под семантикой понимают выражение смысла слов путем их толкования. Однако многие специалисты приходят к выводу о невозможности эффективной алгоритмической реализации семантического анализа через толкования. Таким образом, толкование теряет прикладное значение.

На основе анализа современных семантических методов в целом, а также углубленного исследования семантического аппарата систем ФРАП, ДИАЛИНГ в работе был разработан алгоритм семантического анализа, основанный на использовании Русского общесемантического словаря (РОСС) [16, 23].

Продолжением РОСС являются множество конкретных предметных словарей, а также словарь словосочетаний.

В зависимости от предметной области, к которой относится переводимый текст, одно и то же слово имеет разные значения. Поэтому в специализированных словарях помещается специфическая для каждой предметной области тезаурусная и энциклопедическая информация. Специализированные словари формируют терминологическую основу при переводе соответствующих текстов, хотя, конечно, они не в состоянии объять необъятное и содержать абсолютно всю лексику из всех областей и подобластей данной тематики.

Для того, чтобы автоматически выбрать при переводе один или несколько терминологических словарей необходимо классифицировать тексты, поступающие в систему машинного перевода. Задача классификации документов будет рассмотрена в 5 главе.

В словарь словосочетаний вносятся словосочетания, которые при анализе текста рассматриваются как единый комплекс, то есть приравниваются к однословным единицам.

Каждый из таких словарей организуется в виде отдельной базы данных. В конечном счете, все они должны быть связаны перекрестными ссылками. То есть будут представлять собой систему словарей или "гиперсловарь" системы автоматического анализа текста.

Подробное описание семантического словаря можно найти в работах Леонтьевой Н.Н. [1995, 1997]. В следующем разделе остановимся на свойствах, структуре и словарных статьях словаря, используемых при семантическом анализе русского языка.

4.1. Семантические словари и словарь РОСС

Словарь - это центральное понятие если не всей лингвистики, то уж заведомо любой прикладной системы обработки текста, т.е. лингвистического транслятора. В словаре отражается философия системы, по словарю можно проследить уровни и сам язык описания, принятый в системе, словарная информация позволяет судить о семантической силе системы. Словарь должен примирить языковую теорию и ту практику использования, для которой он предназначен. К настоящему времени можно считать устоявшимся набор характеристик (словарных полей), по которым ведется описание слов в сколько-то развитых системах обработки текста. Так, словари модели Смысл-Текст [17] делают упор на систематизации лексического состава языка и формулировке законов сочетаемости; перевод многих языковых единиц в ранг полусвободных (значений лексических функций от других единиц) позволяет значительно сократить количество традиционно приписываемых им значений и снижает произвол составителей словарных описаний.

Масштабные лексикографические работы, проводимые под руководством Ю.Д.Апресяна и продолжающие традицию модели Смысл-Текст, при установке на исчерпывающее лингвистическое "портретирование" слова концентрируют внимание на тонких смысловых различиях близких по значению или однотипных слов.

В системах, создаваемых под руководством С.Ниренбурга, все большее значение придается онтологическим аспектам описания значений: слово описывается по его месту в Представлении знаний; тем самым обеспечивается связь с предметными областями, из которых наследуются полезные семантические свойства. Дж.Пустейовский развивает генеративный подход к описанию лексики, в рамках которого предполагается собирать и вводить лишь основные типы словарных статей, а описание всего остального массива лексики получать (генерировать) автоматически - применением правил. Такой способ сбора и описания лексики не может не сказаться положительно и на методике и процедурах анализа текста. З.М.Шаляпина в работах по созданию системы машинного перевода с японского языка на русский развивает и углубляет методику компонентного подхода, использующего понятие элементарного смысла: обосновываются связи компонентов значения слова с его валентностями, тем самым вносится стройность в сам язык элементарных смыслов.

РОСС - главный семантический словарь, используемый в систем ПОЛИТекст, ДИАЛИНГ. Он развивает идеи, заложенные в словаре системы ФРАП.

РОСС имеет иерархическую структуру: нижний уровень - поля, принимающие конкретные значения. Верхний уровень составляют зоны - имена групп полей.

РОСС включает семантическое описание по следующему шаблону:

1. Семантический класс лексемы (набор семантических характеристик);
2. Грамматический класс лексемы;
3. Валентная структура лексемы (в терминах семантических отношений);
4. Семантические и грамматические ограничения на выражение каждого актанта из валентной структуры;

В словарной статье словаря РОСС содержится как информация о синтаксической сочетаемости слова, так и информация о семантических характеристиках данного слова, его модели управления, семантических отношениях с другими частями фразы, лексической сочетаемости. Помимо этого имеется большой аппарат для задания адекватного перевода данного слова, или конструкции в которую оно входит, в зависимости от его конкретного употребления. Словарная статья состоит из набора полей, в каждом из которых записывается информация о некотором аспекте поведения слова в языке [16].

На основе описания РОСС был разработан семантический словарь.

4.2. Структура семантического словаря

Словарная статья семантического словаря состоит из набора полей, в каждом из которых записывается информация о некотором аспекте поведения слова в языке.

4.2.1. Формат словарных статей

Входом в словарь считается пара <слово, номер значения>. Для каждого входа составляется словарная статья – набор пар вида <название поля, значение поля>. Между названием поля и его значением ставится знак ‘=’. Название поля состоит из собственно названия и (факультативно) некоторого набора индексов. Например: ГХ1, СХ, СХ1 и т.д. Если поле идет без индекса, значит оно относится к главному слову (обозначается С), если с индексом n – значит к актанта с номером n (обозначается A_n). За каждым названием поля зафиксировано некоторое значение.

Приведем перечень основных полей семантического словаря.

Таблица 4.1.

Поле	Расшифровка	Примеры значений поля
КАТ	Категория лексемы	ЭТК.ОБ, ЭТК.СИТ...
СХ	Семантическая характеристика слова	ОДУШ, ФИН...
СХ1,...,СХ4	Семантические ограничения актантов (1,...,4)	ОДУШ ФИН...
ВАЛ	Валентная структура слова	СУБ, А1, С
ДОП	Дополнительные смысловые отношения между актантами	АДР, А3, А4
НЕСОВМ	Несовместимость способов реализации валентностей	НЕСОВМ, А1, А2
ЛФ	Лексические функции	Oper1: оказывать

Ниже будут описаны те поля и свойства семантического словаря, на основе которых осуществляется семантический анализ русского языка.

4.2.2. Семантические характеристики

Семантические характеристики (СХ) в семантическом словаре играют важнейшую роль в семантическом описании слов. Из СХ строятся формулы (с логическими связками и, или). Каждому слову приписана некоторая формула, составленная из СХ.

Также для каждого слова фиксируется валентная структура $\langle A_1, \dots, A_7 \rangle$, где A_i - описание актанта, которое является парой $\langle \Gamma X_i, C X_i \rangle$, где ΓX_i - некое описание грамматического выражения актанта в предложении, а $C X_i$ - семантическое описание актанта, и $C X_i$ - формула, составленная из СХ.

Под валентностью понимается способность слова сочетаться в тексте с другой языковой единицей, прежде всего с другим словом (ср. термин «валентность» в химии, служащий для описания способности химических элементов образовывать соединения той или иной структуры) [5]. Например, глагол *просить* предполагает, что при нем могут быть указаны проситель (тот, кто просит), предмет просьбы (то, о чем или что просят) и адресат просьбы (тот, кого или у кого просят). Поэтому говорят, что глагол *просить* трехвалентен (кто, кого, о чем). Множество валентностей глагола образует его *валентную структуру*. Валентности, как принято говорить, «заполняются»; заполнители валентностей слова называются его *актантами* [6]. В принципе слово может быть валентно не только на другое слово, но и на словосочетание или даже предложение.

Хотя изначально СХ вводились как простые селективные ограничения, отбраковывающие некоторые связи, проведенные синтаксическим анализом, теперь за каждой из них закреплено определенное значение [16, 22]. В качестве примера, ниже приведем перечень (табл. 4.2.) некоторых семантических характеристик, используемых для семантического анализа (полный список семантических характеристик см. в приложении 1).

Синтаксис записи в поле СХ простой: СХ даются перечислением через запятую.

Пример: статья дом1

СХ = 1 АБСТР

ВМЕСТЛ

Таблица 4.2.

СХ	Комментарий	Примеры слов с таким СХ
АБСТР	Любое абстрактное существительное или прилагательное	модель план, тенденция, обстоятельство
АРТ	Артефакт. Все, что сделано человеком	машина, хлеб, памятник
ВЕЩВО	Любое название химического вещества или того, что можно как-либо дозировать, отмерять, продавать по весу или объему	аммиак, бензин
ВМЕСТЛ	Все, что предназначено для содержания чего-либо другого	мешок, гараж
ВРЕД	Все, к чему человек обычно относится как к нежелательному.	катастрофа, война.
ГЕОГР	Любой географический объект	остров, река.
ГОС	Любое название государства или тип государства	республика
ДВИЖ	Глаголы движения	идти, бежать
ДОЛЖ	Должность, профессия, социальный статус	повар, врач
Д-УСТР	Деталь устройства	валик

4.2.3. Общая категоризация лексики (поле КАТ)

КАТ - семантическая категория входной единицы, она определяется способом отображения в семантической структуре (или информационном языке-посреднике - ИЯП).

Элементарное высказывание на ИЯП имеет вид $R(A,B)$, где R - отношение, A и B - термы. Например: ПРИЧИНА (взрыв АЭС, катастрофа); ЛОКАЛИЗАЦИЯ (Чернобыль, АЭС).

Поле КАТ принимает значения ЭТК, ОТН, ОПЕР.

Категория задает верхний уровень семантической классификации всей лексики:

1. ЭТК - слова-этикетки, занимают позицию A или B в формуле $R(A,B)$. Слова этой категории - самый большой, открытый и подвижный класс слов. Слова категории ЭТК образуют три подкласса: признаки, ситуации и объекты (ПРИЗН, СИТ и ОБ).

В очевидных случаях слову нужно приписывать уточненную категорию: ЭТК-ПРИЗН (практически всем прилагательным и наречиям), ЭТК-СИТ (например, словам "война", "обсуждение", "утечка" и др.) или ЭТК-ОБ (слова "президент", "статья", "законопроект"). В неоднозначных или неясных случаях ("система", "режим", "право") приписывается общая категория (просто ЭТК).

Дальнейшая семантическая дифференциация слов-этикеток задается значениями полей СХ (семантическая характеристика) и ВАЛ (смысловые валентности).

2. ОТН - слова-отношения (или смысловые отношения - СО), занимают позицию R в формуле $R(A,B)$, например, РАВНО(A,B), АДРЕСАТ(A,B).
3. Слова с категорией ОПЕР не имеют собственного смысла, а лишь модифицируют уже существующее семантическое пространство. Например, слова *не*, *еще*, *уже* и т.д.

4.2.4. Семантическое отношение (поле ВАЛ)

Семантическое отношение - это некая универсальная связь, усматриваемая носителем языка в тексте. Эта связь бинарна, т.е. она идет от одного семантического узла к другому узлу.

Поле ВАЛ перечисляет набор валентностей слова. В отличие от СХ, имена смысловых валентностей становятся именами дуг семантического представления (семантического дерева) и поэтому являются в принятой системе семантического анализа более важной характеристикой, чем СХ.

Формат записи семантического отношения следующий:

$R(A,B)$, где R – название семантического отношения, A – зависимый член отношения, B – управляющий член отношения.

Для конкретных A,B и отношения R направление выбирается таким образом, чтобы формула $R(A,B)$ была эквивалентна утверждению, что "A является R для B". Соответственно, формула $R(B,A)$ должна быть эквивалентна утверждению "B является R для A". Семантическое отношение формирует и организует текст.

Среди семантических отношений достаточно много таких, которые сейчас уже повсеместно считаются универсальными. Семантические отношения, используемые в системе, заимствованы у систем машинного перевода ФРАП[12] и ДИАЛИНГ[23].

В качестве примера, в таблице 4.3. приведены некоторые из семантических отношений (полный список семантических отношений см. в приложении 2).

Таблица 4.3.

ВАЛ	Комментарий	Примеры
АВТОР	Обычно одушевленный участник, в результате действий которого появляется что-то (ОБ.РЕЗЛТ например)	Роман Толстого
АГЕНТ	Одушевленный участник, контролирующей ситуацию. В отличие от СУБ у АГЕНТа есть цель и про него можно сказать "сделал что-то с целью..."	Мы сократили отставание
АДР	Адресат; тот на кого направлено действие, или для кого оно совершается	Я отдал стул отцу
ЗНАЧ	значение (какого-либо параметра)	Высота дома – 20 метров
ИДЕНТ	Идентификатор (название чего-либо). Валентность заполняется некоторым номером, стоящим справа, либо примыкающим определением. Не может заполняться группой ФИО и именами-фамилиями. пр. : абзац	Высота дома – 20 метров
ИМЯ	Заполняется только группой ФИО или одиночными фамилиями-именами	Дворник Степанов
ИНСТР	Инструмент	Резать ножом

В принципе количество смысловых валентностей, приписываемых слову, не ограничивается, и порядок их записи не важен. В настоящей версии БД семантического анализа решено вводить при слове до 4 валентностей.

4.2.5. Поле ДОП, НЕСОВМ

Поле ВАЛ естественным образом дополняется информацией о том, какими смысловыми отношениями могут быть связаны между собой объявленные актанты. Информация фиксируется в поле ДОП (дополнительные смысловые отношения между актантами, отношения нумеруются и отделяются знаком ";"):

Пример статья ЗГЛ = компенсировать;

ВАЛ = АКТ,А1,С; ПРИЧ,А2,С; АДР,А3,С; СОДЕРЖ,А4,С

ДОП = 1. АКТ,А1,А4; 2. ПАЦИЕН,А3,А2; 3. АДР,А3,А4

Тем самым в поле ДОП перечисляются (формально задаются) те ситуации, которые можно ожидать в развертывании текста, если реализовано значение слова «компенсировать».

НЕСОВМ - несовместимость способов реализации валентностей. Формат записи поля НЕСОВМ следующий:

НЕСОВМ = <номер актанта> , <номер актанта>.

В поле записываются актанты слова, одновременное выражение которых при данном слове невозможно.

Пример: статья сблизать¹

ВАЛ = СУБ, А1 , С

АКТ, А2 , С

П-АКТ, А3 , С

В-АКТ, А4 , С

НЕСОВМ = А2, А3

А2, А4

4.2.6. Поле ЛФ, ЛХ_і

1. В поле ЛФ записываются лексические функции данного слова.

Пример: статья давление¹

ЛФ = Орег¹: оказывать

Орег²: испытывать

Орег³: подвергаться

2. В поле ЛХ_і (лексическая характеристика) записываются слова, которые с большой вероятностью могут заполнять *i*-ую валентность.

Пример: статья сковывать¹

ГХ₁ = 1 подл : И

ЛХ₁ = страх

Лед

ЛХ_і употребляется тогда, когда в позиции *i*-го актанта могут стоять слова из ограниченного набора, и придумывать для них СХ не имеет смысла.

4.3. Построение семантического представления предложения

4.3.1. Введение

Традиционный школьный синтаксис, который строится на понятии согласования, управления и примыкания, позволяет очертить круг синтаксических отношений. Все алгоритмы, которые ищут эти отношения и их композиции, можно назвать синтаксисом. Семантикой мы называем те алгоритмы, которые, используя смысл слов и выражений, устанавливают отношения, которые не вычисляются напрямую из синтаксических отношений.

В ходе семантического анализа текста строится его семантическое представление (СемП), которое представляет собою связный ориентированный граф, состоящий из семантических узлов и отношений между ними.

Семантические отношения уже были описаны выше. Семантический узел – одно из центральных понятий семантической теории. Определение узла формулируется так [22]:

Семантический узел – это такой объект текстовой семантики, у которого заполнены все валентности.

Как и на всех этапах анализа, семантические узлы образуются из слов исходного предложения. Главные источники гипотез о составе семантического узла, семантической структуры предложения дают синтаксический анализ, семантический словарь, терминологические словари, словарь словосочетаний. В текущей версии программы семантического анализа семантическое представление предложения строится на основе анализа словарных статей семантического словаря.

Построение семантического представления фрагмента текста состоит из следующих этапов (рис.4.1.).

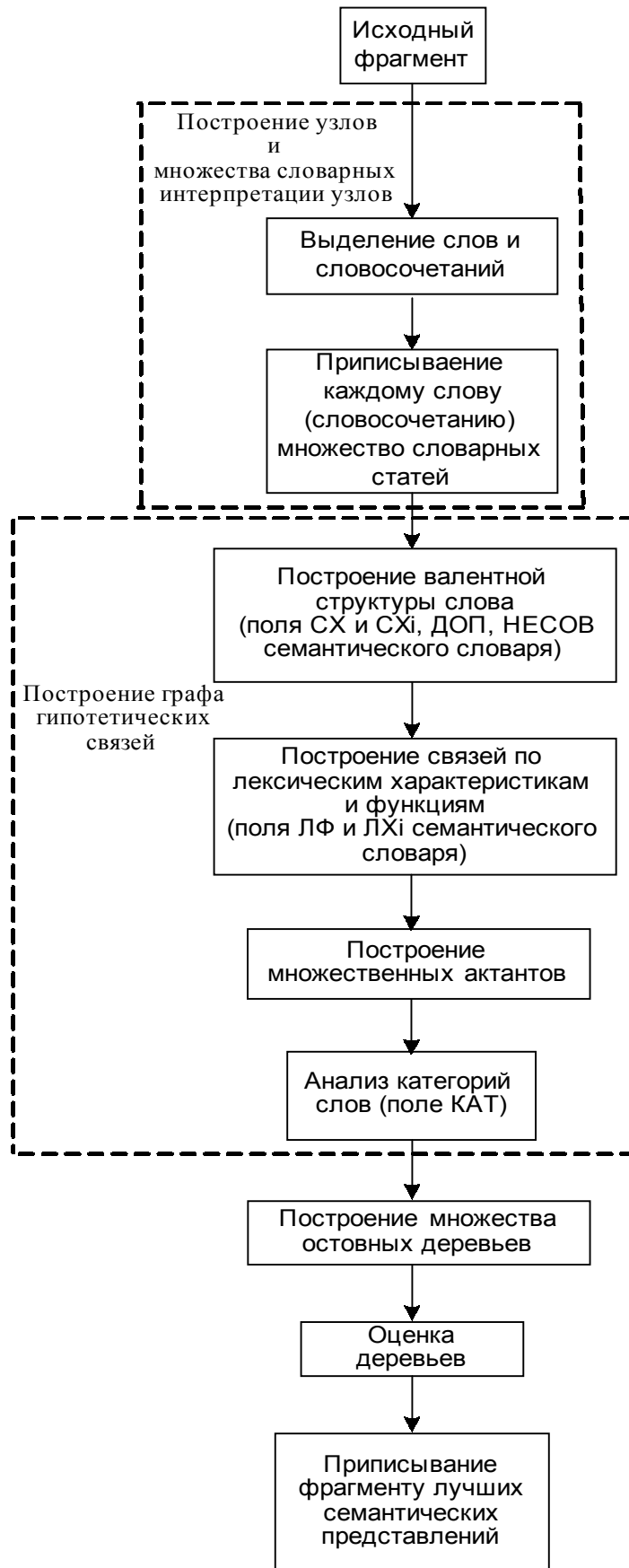


рис.4.1. Основные этапы построения семантического представления фрагмента текста

4.3.2. Построение узлов и множества словарных интерпретации узлов

Первый этап - построение узлов и словарных интерпретаций каждого узла. В большинстве случаев узел - это одно слово, но могут быть и словосочетания (например, «застать врасплох»). Словарная интерпретация узла - это множество словарных статей, в каждой из которых записано, какими другими узлами может управлять данный узел.

Процедура построения словарных интерпретаций приписывает каждому семантическому узлу множество словарных входов, к которым может быть приравнен этот семантический узел. Например, слову “стекло” будет приписано множество словарных входов {стекло1, стекло2}.

4.3.3. Построения графа гипотетических связей

На втором этапе первичного семантического анализа строится граф гипотетических связей. Граф состоит из узлов и семантических отношений между ними.

Для каждого узла графа проводятся все связи, которые можно провести к другим узлам, основываясь на словарных статьях узлов.

Граф гипотетических связей почти всегда содержит много лишних гипотез, которые придется отбрасывать на последнем этапе.

Построение валентной структуры слова

Когда алгоритм пытается собрать валентную структуру слова, критерием установления связи служит соответствие значения полей CX приписанных актанту, значению полей CX_i , приписанных главному слову.

Установление связи по семантическим характеристикам осуществляется следующим образом. Пусть *Host* - словарная статья узла, от которого идет рассматриваемое отношение (далее просто отношение).

Пусть *Slave* – словарная статья узла, к которому идет отношение. Пусть отношение, которое мы проверяем, стоит *i*-ым по порядку в словарной статье слова *Host*. Поля CX содержат дизъюнкцию конъюнкций семантических характеристик или отношений. Дизъюнкты нумеруются начиная с 1.

Например:

$$CX(\text{Slave}) = \begin{array}{l} 1 \text{ ФИН, ОРГ} \\ 2 \text{ ОДУШ} \end{array}$$

Считается, что отношение удовлетворяет критерию семантических характеристик, если какой-либо из дизъюнктов $CX_i(\text{Host})$ полностью вкладывается в какой-либо дизъюнкт $CX(\text{Slave})$.

Например:

$CX_i(\text{Host}) = \text{ОРГ}$

$CX(\text{Slave}) = \text{ОРГ, ФИН}$

Кроме этого, при построении валентной структуры слова добавляются валентности из так называемых добавочных статей, устанавливающие отношения между актантами слова (поля ДОП, НЕСОВ).

Построение связей по лексическим функциям и характеристикам слов (поля ЛФ, ЛХ_i)

Пусть в словаре для слова X есть запись $ЛФ = F : Y$, где F – лексическая функция-параметр, а Y – слово. Значения лексических функций-параметров записывается в словарной статье слова-ситуации в словаре (поле ЛФ). Необходимо найти все такие пары X, Y в пределах одного фрагмента и поместить их в специальное множество. В зависимости от типа лексических функций для пары X, Y должны выполняться синтаксические ограничения:

- если F – функция $Орег$, то Y может быть вторым актантом X ;
- если F – функция $Func$, то Y может быть первым актантом X .

Если пара X, Y удовлетворяет всем вышперечисленным критериям, то она помещается во множество найденных лексических функций. Лексические функции отображаются в семантическом представлении.

Также проводится анализ по лексическим характеристикам слова (поле ЛХ_i). Как было сказано выше, в поле ЛХ_i (лексическая характеристика) записываются слова, которые с большой вероятностью могут заполнять i -ую валентность.

ЛХ_i употребляется тогда, когда в позиции i -го актанта могут стоять слова из ограниченного набора, и придумывать для них CX не имеет смысла.

ЛХ_i = <слово> означает, что <слово> может быть i -м актантом входного слова.

Построение множественных актантов

Множественный актант возникает там, где одна валентность предиката заполняется многими актантами. Синтаксический однородный ряд всегда переходит во множественный актант.

Например:

Петя и Маша => МНА_n(*Петя, Маша*).

Таким образом, множественный актант – это множество актантов, заполняющих одну валентность предиката и упорядоченных между собой оператором однородности.

Основной механизм работы с множественными актантами следующий. Сначала по стандартным законам проводятся связи к членам однородного ряда от потенциального хозяина и от оператора однородности. Затем, запускается процедура, которая проходит по всем операторам однородности, если X подчиняется одновременно оператору однородности и другому узлу Y, тогда проводится стрелка от Y к оператору однородности (см.рис.4.2.).

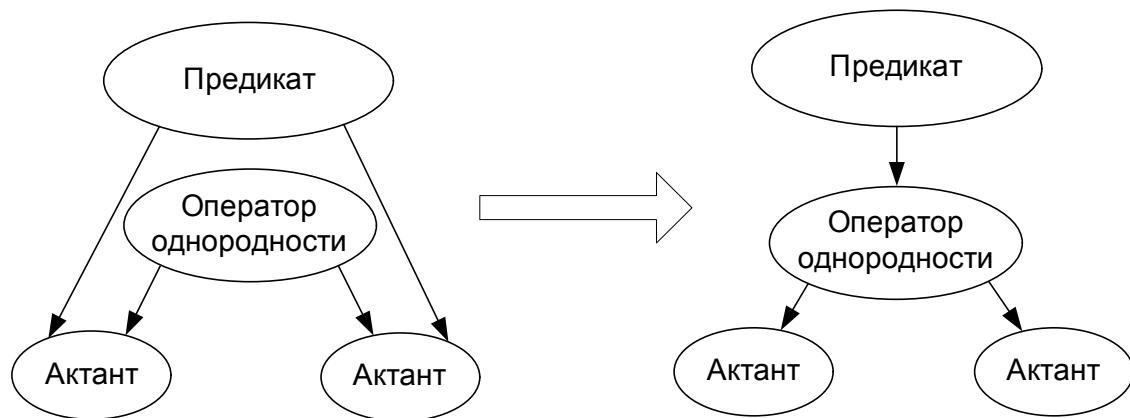


рис.4.2. Построение множественных актантов

Анализ категорий слов (поле КАТ)

Разбиение лексики на категории, информация о которых заносится в поле КАТ, производится в зависимости от того, какое место в формуле R(A,B) занимает рассматриваемая лексическая единица с точки зрения ее семантики. Поле КАТ принимает значения ЭТК, ОТН, ОПЕР.

Слова-этикетки занимают в семантической формуле R (A, B) позицию одного из термов A или B, то есть представляют собой семантический узел.

Слова - отношения занимают в формуле R (A, B) позицию R.

Существует еще одна категория, слова которой не являются ни семантическими узлами, ни отношениями – это категория ОПЕР. Семантику этих слов нельзя описать при помощи формулы R(A, B), так как в семантическом представлении словам этой категории

не сопоставляются ни узлы, ни связывающие их отношения. Значение этих слов-операторов накладывается на уже построенный семантическое представление, и преобразует его семантику. Семантическое поведение слов этой группы чрезвычайно индивидуально. Оно описывается специальной грамматикой.

Часто помета ОПЕР встречается как дополнительная, то есть она может появляться в поле КАТ у слов, которые уже отнесены к некоторой другой категории. Это означает, что в семантике их присутствует компонент, не позволяющий считать их исключительно семантическим узлом или семантическим отношением. На практике это означает, что семантическое поведение таких слов нетривиально и его удобнее описывать не общими правилами грамматики, а индивидуальными правилами [23].

4.3.3. Построение и оценка древесных вариантов

Процедура построения графа гипотетических связей почти всегда строит граф, отношений в котором больше, чем нужно.

Для построения вариантов деревьев по графу семантических отношений и выбора лучшего дерева (оценка деревьев) можно использовать подход, принятый в системе ДИАЛИНГ [23].

Подход заключается в следующем. По графу гипотетических связей строится множество остовных деревьев, в которых нет двух семантических отношений с одинаковым непустым названием, идущих от одного узла к двум разным узлам.

Остовное дерево связного графа – это максимальный подграф, который является деревом.

Процедура построения вариантов деревьев работает в несколько этапов. Сначала, строится множество всех максимальных подграфов, в которых в каждую вершину входит не более одной стрелки (множество подграфов S_1). Затем из каждого подграфа из множества S_1 выделяются следующие подграфы, в которых из всех циклов удалено по какой-нибудь стрелке (множество S_2). Таким образом, получается множество ациклических графов, в каждый узел которого входит не более одной стрелки. По определению, если такие графы связаны, то они являются деревьями. Затем из каждого подграфа множества S_2 выделяются следующие подграфы, в которых из одного узла не выходит несовместимых отношений (множество S_3). Несовместимость может быть общая и словарная. Общая несовместимость – это запрещение двум одноименным стрелкам выходить из одного узла. Словарная несовместимость – это те конкретные правила несовместимости валентностей, которые записаны в словарной статье данного узла. После уничтожения несовместимых отношений из множества подграфов выбираются те, которые имеют максимальное число компонент связностей.

При оценке деревьев для каждого дерева вычисляется его вес. Остовные деревья оцениваются по следующим критериям.

1) Основные критерии

- одна валентность не может заполняться дважды;

2) Дополнительные критерии

(структурные)

- Проективность;
- Длина отношений;
- Порядок актантов в тексте;
- Нарушение оператора МНА;
- Согласование по СХ;

(словарные)

- Общее число валентностей, которые заполнены одним из значений стандартной лексической функции (поле ЛХ);
- Число узлов, построенных на лексических функциях;

(грамматические)

- Равна или нет вершина построенного графа сказуемому в синтаксическом анализе;
- Число узлов, нарушающих грамматические ограничения;
- Удовлетворяет ли корень дерева морфологическим критериям;
- Проверка согласования подлежащего и сказуемого;

Чем больше вес остовного дерева, тем хуже само дерево.

Для каждого фрагмента строится множество лучших семантических представлений, которое является результатом работы первично семантического анализа.

4.3.4. Основные принципы работы программы семантического анализа

Текущая версия программы семантического анализа представляет собой исследовательскую (экспериментальную?) разработку и осуществляет упрощенный семантический анализ простых предложений русского языка.

На вход семантического анализа подается одно предложение на русском языке. На выходе семантический анализ строит семантическое представление (СемП) предложения, представленное в виде списка семантических отношений между узлами предложения. При этом выполняются следующие действия:

1. Разбиение предложения на слова.
2. Лемматизация слов входного предложения (приведение текстовых форм слова к словарным);
3. Поиск полученных лемм в словаре;
4. Получение для каждого узла (слова) его словарной интерпретации;
5. Построение всех возможных гипотетических связей по СХ и СХ_i;
6. Получение перечня валентностей (семантических отношений) для каждого узла;
7. Вывод списка построенных отношений.

Формат семантического отношения следующий:

$R(A,B)$, где R – название семантического отношения, A – зависимый член отношения, B – управляющий член отношения.

Например, для предложения «Он строил дом» будет построен следующий список отношений.

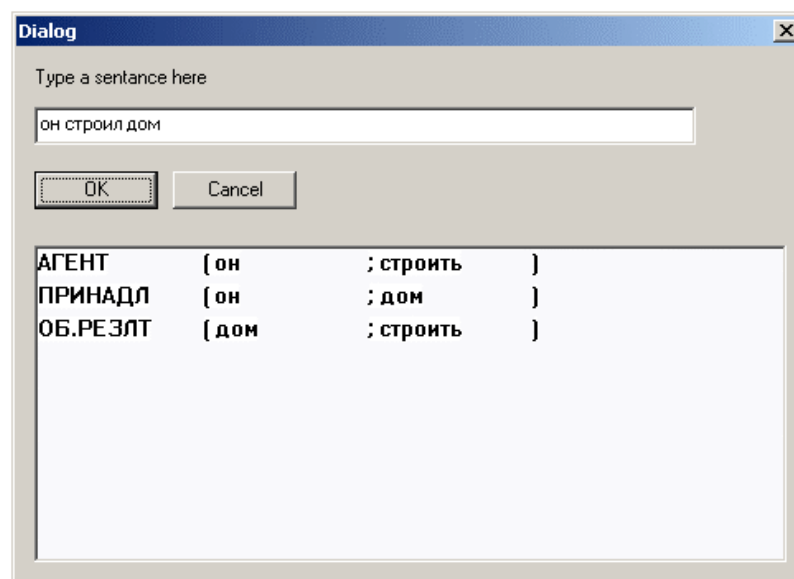


рис. 4.3. Пример анализа предложения

Теоретически, главные источники гипотез о составе семантического узла, семантической структуры предложения дают синтаксический анализ, семантический словарь, терминологические словари, словарь словосочетаний. В текущей версии программы семантического анализа семантическое представление предложения строится на основе анализа словарных статей семантического словаря. Поэтому на рис.4.3. построенных отношений больше чем нужно (лишнее отношение ПРИНАДЛ(дом, он)). Кроме этого, необходимо оценивать построенное СемП.

Про построении СемП предложения в виде графа:

- название семантического отношения R становится именем дуги СемП
- управляющий член отношения В переходит в семантический узел, соответствующий главному слову,
- зависимый член отношения А переходит в семантический узел, из которого в СемП выходит стрелка, входящая в В.

Добавление слов и их словарных статей осуществляется с помощью диалогового окна (рис.4.4.).

рис.4.4. Ввод леммы и ее словарной статьи

В настоящей версии словаря решено вводить до 3 семантических ограничений актантов (поля СХ1...СХ3).

Программа семантического анализа написана на Microsoft VC++ 6.0.

Выводы

На основе анализа современных семантических методов в целом, а также углубленного исследования семантического аппарата систем ФРАП, ДИАЛИНГ разработан алгоритм семантического анализа, основанный на использовании Русского общесемантического словаря (РОСС).

На основе описания РОСС разработан семантический словарь. Рассмотрены свойства, структуры словарных статей семантического словаря, используемых при семантическом анализе русского языка.

Главные источники гипотез о составе семантического узла, семантической структуры предложения дают синтаксический анализ, семантический словарь, терминологические словари.

В текущей версии программы семантического анализа семантическая структура предложения строится на основе анализа словарных статей семантического словаря. Семантический анализ строит семантическое представление одного предложения на русском языке.

5. Классификация текстов

В зависимости от предметной области, к которой относится переводимый текст, одно и то же слово имеет разные значения. Специализированные словари формируют терминологическую основу при переводе соответствующих текстов, хотя, конечно, они не в состоянии объять необъятное и содержать абсолютно всю лексику из всех областей и подобластей данной тематики.

Для того, чтобы автоматически выбрать при переводе один или несколько терминологических словарей необходимо классифицировать тексты, поступающие в систему машинного перевода.

Целью задачи классификации является определение для каждого документа одной или нескольких из заранее заданных категорий, к которым этот документ относится. Особенностью задачи классификации является предположение, что множество классифицируемых документов не содержит ``мусора'', т.е. каждый из документов соответствует какой-нибудь из заданных категорий.

Частным случаем задачи классификации является задача *тематической классификации*. Здесь каждая категория - это некоторая тематика, а цель классификации - определить тематику документа.

Мы рассматриваем задачу тематической классификации документов.

Большинство существующих подходов к анализу текстов можно разбить на два класса. К первому классу относятся простые, быстрые, но грубые механизмы анализа; чаще всего это подходы, использующие формальные статистические методы, основанные на частоте появления в тексте слов различных тематик. Второй класс формируют достаточно изощренные, дающие хороший результат, но сравнительно медленные подходы, основанные на лингвистических методах.

Эффективным же можно считать такой метод, который сочетал бы в себе «простоту» статистических алгоритмов с достаточно высоким качеством обработки лингвистических методов.

В тоже время, как показала практика, для достижения приемлемого качества решения практических задач компьютерного анализа текстовой информации (автоматическое аннотирование, тематическая категоризация и т.д) не требуется полный грамматический анализ фразы. Достаточно выделить наиболее информативные единицы текста - ключевые слова, словосочетания, предложения и фрагменты, причем в качестве характеристики информативности хорошо работает частота повторения слов в тексте [8, 9].

Поэтому в основе предлагаемой модели лежит представление смысла текста в форме сети, узлы которой представлены множеством часто встречающихся понятий текста - слов и устойчивых словосочетаний, из числа которых исключены общепотребимые слова. Узлы сети связаны между собой с различной силой связи, причем сила связи коррелирована с частотой совместной встречаемости понятий в предложениях текста.

Сеть может быть автоматически построена на базе множества текстов. Ее можно рассматривать как модель предметной области или как модель некоторой тематики. Автоматически построенные модели заданных тематик могут быть использованы для анализа неизвестных документов.

Таким образом, задачу тематической классификации текстов можно сформулировать следующим образом: построение моделей тематик, построение описания (модели) рассматриваемого документа и оценка близости между описаниями тематик и описанием документа.

5.1. Построение модели сети

Сеть есть набор элементов, представляющих понятия предметной области (ключевые слова и словосочетания), которые связаны между собой с различной силой связи и может быть описана матрицей весов связей:

$$W = [w_{ij}] \quad (5.1)$$

Пусть имеется модель сети (1), представленная в виде

$$P^N = [p(j|i)], i=1..N, j=1..N \quad (5.2)$$

где $p(j|i) \sim w_{ij}$ – условная вероятность появления j -го понятия в связи с i -м, а N – количество элементов сети.

Обратимся к задаче построения модели такой сети.

5.1.1. Выделение ключевых слов

Выбор решения

В качестве решения возможно использование тематических словарей, составленных «вручную», однако имеется одна проблема: изменчивость языка. Язык постоянно меняется: появляются новые слова, термины, фразеологизмы, другие лексические единицы устаревают, выходят из употребления или меняют значение. Поэтому (хотя бы частичная) автоматизация процесса составления словарей (сети) представляется весьма полезным делом.

Для выделения понятий сети, представляющих слова и связанные словосочетания, может быть применен статистический алгоритм [18], основанный на анализе частоты встречаемости цепочек слов различной длины и их вхождения друг в друга.

Во всех созданных человеком текстах можно выделить статистические закономерности. Никому не удастся обойти их. Кто бы их ни писал, какой бы язык он при этом ни использовал, внутренняя структура текста останется неизменной. Она описывается законами Дж. Зипфа (*George K. Zipf*). Законы Зипфа универсальны. Зипф предположил, что природная лень человеческая ведёт к тому, что слова с большим количеством букв встречаются в тексте реже коротких слов. Основываясь на этом постулате, Зипф вывел два универсальных закона.

Законы Зипфа

Первый закон Зипфа "ранг -- частота"

Если измерить количество вхождений каждого слова в текст и взять только одно значение из каждой группы, имеющей одинаковую частоту, расположить частоты по мере их убывания и пронумеровать (порядковый номер частоты называется рангом частоты), то наиболее часто встречающиеся слова будут иметь ранг 1, следующие за ними - 2 и т.д. Вероятность встретить произвольно выбранное слово будет равна отношению количества вхождений этого слова к общему числу слов в тексте.

Вероятность = Количество вхождений слова / Число слов

Зипф обнаружил следующую закономерность: произведение вероятности обнаружения слова в тексте на ранг частоты – константа (С).

$C = (\text{Количество вхождений слова} \times \text{Ранг частоты}) / \text{Число слов}$

Это функция типа $y=k/x$ и её график - равнобочная гиперболa. Следовательно, по первому закону Зипфа, если самое распространенное слово встречается в тексте, например, 100 раз, то следующее по частоте слово с высокой долей вероятности, окажется на уровне 50.

Значение константы С в разных языках различно, но внутри одной языковой группы остается неизменно, какой бы текст мы ни взяли. Так, например, для английских текстов константа Зипфа равна приблизительно 0,1.

Второй закон Зипфа "количество -- частота"

В первом законе не учтён тот факт что, разные слова могут входить в текст с одинаковой частотой. Зипф установил, что частота и количество слов, входящих в текст с этой частотой, тоже связаны между собой. Если построить график, отложив по одной оси (оси X) частоту вхождения слова, а по другой (оси Y) - количество слов в данной частоте, то получившаяся кривая будет сохранять свои параметры для всех без исключения созданных человеком текстов. Как и в предыдущем случае, это утверждение верно в пределах одного языка. Однако и межъязыковые различия невелики. На каком бы языке текст ни был написан, форма кривой Зипфа останется неизменной (рис.5.1). Могут немного отличаться лишь коэффициенты, отвечающие за наклон кривой (в логарифмическом масштабе, за исключением нескольких начальных точек, график - прямая линия).

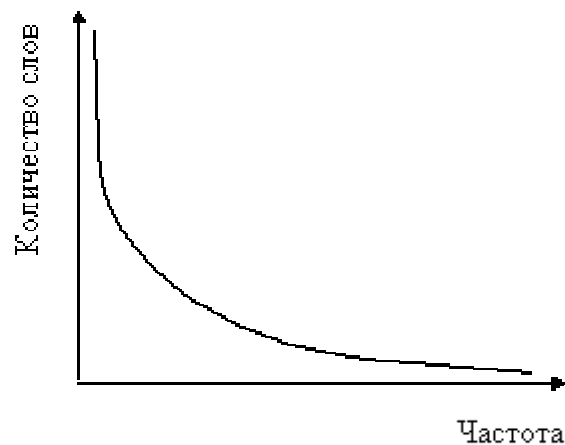


рис 5.1. График зависимости частоты вхождения слова от количества слов в данной частоте

Законы Зипфа универсальны. В принципе, они применимы не только к текстам. В аналогичную форму выливается, например, зависимость количества городов от числа проживающих в них жителей. Характеристики популярности узлов в сети Интернет - тоже отвечают законам Зипфа. Не исключено, что в законах отражается "человеческое" происхождение объекта. Так, например, ученые давно бьются над расшифровкой манускриптов Войнич. Никто не знает, на каком языке написаны тексты и тексты ли это вообще. Однако исследование манускриптов на соответствие законам Зипфа доказало: это созданные человеком тексты. Графики для манускриптов Войнич точно повторили графики для текстов на известных языках.

Воспользуемся законами Зипфа для извлечения из текста слов, отражающих его смысл (ключевых слов).

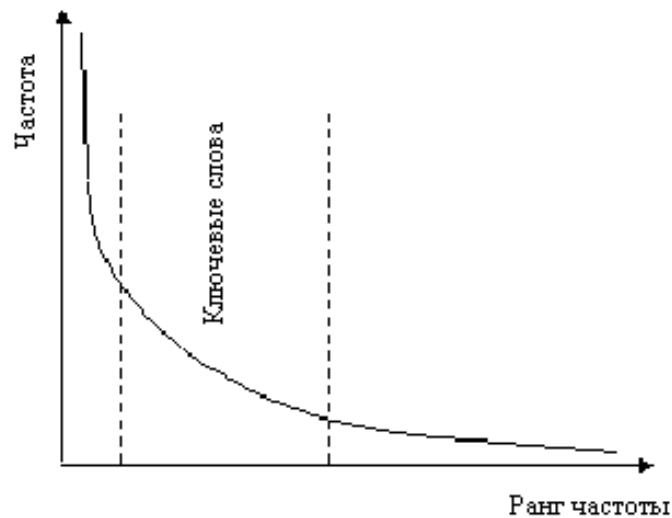


рис 5.2. График зависимости ранга частоты от частоты вхождения слова

Исследования показывают, что наиболее значимые слова лежат в средней части диаграммы (рис. 5.2) Это и понятно. Слова, которые попадаются слишком часто, в основном оказываются предлогами, местоимениями, в английском - артиклями и т.п. Редко встречающиеся слова тоже, в большинстве случаев, не имеют решающего смыслового значения.

От того, как будет выставлен диапазон значимых слов, зависит многое. Если поставить широко – то в ключевые слова будут попадать вспомогательные слова; если установить узкий диапазон – то можно потерять смысловые термины.

Сделать выделение наиболее значимых слов качественнее помогает предварительное исключение исследуемого текста некоторых слов, которые априори не могут являться значимыми и, поэтому являются «шумом». Такие слова называются нейтральными или стоповыми (стоп-словами). Словарь стоп-слов называют стоп-листом. Например, для английского текста стоп-словами станут термины: *the, a, an, in, to, of, and, that...* и так далее.

Весовые коэффициенты

До сих пор рассматривался лишь отдельно взятый документ, не принимая во внимание, что он входит в базу данных наряду с множеством других документов. Если представить всю базу данных как единый документ, к ней можно будет применить те же законы, что и к единичному документу.

Обычно, чтобы избавиться от лишних слов и в тоже время поднять рейтинг значимых слов, вводят инверсную частоту термина. Значение этого параметра тем меньше, чем чаще слово встречается в документах базы данных. Вычисляют его по формуле (5.3).

Инверсная частота термина $i = \log (\text{количество документов в базе данных} / \text{количество документов с термином } i)$ (5.3)

Теперь каждому термину можно присвоить весовой коэффициент, отражающий его значимость:

Вес термина i в документе $j = (\text{частота термина } i \text{ в документе } j) \times (\text{инверсная частота термина } i)$ (5.4)

Таким образом, если рассматривать задачу вычисления веса термина для множества текстов данной тематики (вес термина в модели семантической сети), то, используя (5.3) и (5.4), получаем:

Вес термина i для данной тематики $K_j = (\text{частота термина } i \text{ в тематике } K_j) \times (\text{инверсная частота термина } i)$, где

Инверсная частота термина $i = \log (\text{количество документов (данной тематики) в базе данных} / \text{количество документов (данной тематики) с термином } i)$

Т.о. термин, получивший нулевой или близкий к нулю вес, можно исключить из модели семантической сети.

В качестве терминов могут выступать не только отдельные слова, но и словосочетания.

5.1.2. Оценка параметров сети

Оценка параметров модели сети в форме (5.2) требует определения понятий, а также условных вероятностей $p(j|i)$ появления пары понятий в связи. Провести такую оценку возможно на основе анализа множества текстов, порожденных моделью - эталонных текстов из одного класса в задаче классификации.

Покажем, как провести оценку весов связей.

По определению условной вероятности

$$p(j|i) = p(ij) / p(i) \quad (5.5)$$

где $p(ij)$ - вероятность появления пары понятий в связи, а $p(i)$ - собственная вероятность появления i -го понятия в тексте.

Обозначим набор понятий как вектор $W = (w_i)$, где

$w_i = 1$, если i - ое понятие сети присутствует в наборе

$w_i = 0$ - в противном случае

Тогда $|W| = \sum_i w_i(t)$ есть количество понятий в наборе. Пустой набор будем обозначать W_0 .

Представим предложение как набор входящих в него понятий $W(t) = (w_i(t))$, где $t=1..T$ - порядковый номер предложения в тексте.

Будем считать, что каждое предложение имеет одно порождающее понятие – тему, которое обуславливает появление всех остальных понятий, связанных с ним, но попарно независимых.

В качестве критерия возможной связности понятий используем факт их появления в одном предложении текста. Отсутствие априорной информации на этапе построения модели не позволяет учесть сверхфразовые связи, вследствие чего разумно предположить все понятий равновероятными в качестве тем. Тогда, считая, что каждое из понятий равновероятно связано с любым из других, имеем

$$p(ij|W(t)) = w_i(t) w_j(t) / [\sum_j w_j(t)-1] \text{ для } i \neq j \quad (5.6)$$

$$p(ii|W(t)) \equiv 1;$$

Полная вероятность связи понятий определяется по всему тексту как

$$p(ij) = \sum_t p(ij|W(t)) P(W(t)), t=1..T \quad (5.7)$$

Собственная вероятность появления понятия

$$p(i) = \sum_t p(i|W(t)) P(W(t)) = \sum_t w_i(t) / T, t=1..T \quad (5.8)$$

Окончательно, с учетом (5.7), (5.8) и (5.6) получаем из (5.5) искомую оценку

$$p(j|i) = \sum_t p(ij|W(t)) / \sum_t w_i(t) = \sum_t [w_i(t) w_j(t) / [\sum_j w_j(t)-1]] / \sum_t w_i(t) \quad (5.9)$$

Как видно, выражение в знаменателе представляет собственную частоту встречаемости понятия в тексте (исключая повторы в одном предложении), а выражение в числителе есть частота совместной встречаемости понятий в предложениях текста, нормированная с учетом количества понятий по каждому из предложений.

Таким образом, связь от понятия i к понятию j предлагается характеризовать весом w_{ij} , который в простейшем случае определяется как

$$w_{ij} = f_{ij} / f_i,$$

где f_{ij} – частота совместной встречаемости понятий в предложениях текста, а f_i – собственная частота встречаемости понятия в тексте. Как видно, вес связи отражает условную вероятность того, что при упоминании в тексте понятия i речь также идет о понятии j .

Для уточнения модели можно учесть, что некоторые связи не наблюдаются явно в предложениях текста, однако подразумеваются автором. Их скрытое влияние выражается в том, что вместо $p(j|i)$ правильнее было бы использовать вероятность $p(j|q)p(q|i)$, где q – ненаблюдаемое понятие. С учетом этого взамен $p(j|i)$ можно использовать уточненную оценку $\tilde{p}(j|i)$, учитывающую связь через третьи понятия:

$$\tilde{p}(j|i) = \max_q \{ p(j|q)p(q|i) \}, q=1..N \quad (5.10)$$

На практике при использовании модели из предложений текста следует исключить общеупотребимые слова.

5.2. Порождение текста на основе сети

Покажем, как оценить вероятность того, что произвольный текст был порожден на основе заданной модели.

Как было сказано выше, считаем, что каждое предложение имеет одно порождающее понятие – тему, которое обуславливает появление всех остальных понятий, связанных с ним, но попарно независимых.

Тогда вероятность порождения предложения $W(t)$ от понятия-темы с номером m можно определить как

$$P(W(t) | m) = \prod_i p(i | m)^{w_i(t)}, \quad i=1..N, \quad \text{где } m \text{ – порождающее понятие} \quad (5.11)$$

Учитывая то, что порождающее понятие-тема достоверно неизвестно, и полагая его появление обусловленным понятиям предшествующего предложения, с привлечением формулы полной вероятности представим вероятность порождения предложения как условную:

$$P(W(t) | W(t-1)) = \sum_m w_m(t) P(W(t) | m) P(m | W(t-1)), \quad m=1..N \quad (5.12)$$

Полагая равновероятным, что любое из понятий предложения $W(t-1)$ могло обусловить тему предложения $W(t)$, имеем:

$$P(m | W(t-1)) = \sum_j w_j(t-1) p(m | j) / \sum_j w_j(t-1), \quad j=1..N \quad (5.13)$$

Уравнение (12) описывает процесс порождения текста как марковский процесс первого порядка.

Для формальной корректности модели следует положить, что

$$W(0) = W_0;$$

$$P(W_0) = P(W_0 | W(t)) = 1/2^N; \quad (5.14)$$

$$p(m | W_0) = 1/w_j(t)$$

Таким образом, порождающее предложение текста считается пустым, появление пустого предложения считается необусловленным и необуславливающим событием, а в предложении, порожденным пустым, все понятия полагаются равновероятными в качестве тем. Появление пустого предложения (в графическом тексте аналогичного началу нового абзаца) означает возможность переключения внимания автора сообщения на новую тему, выбор которой обусловлен факторами, не поддающимися оценке в рамках принятой модели.

Полная вероятность порождения наблюдаемого текста моделью есть вероятность соответствующей реализации марковского процесса (5.12), вычисляемая с учетом (5.14) как

$$P^* = \prod_t P(W(t) | W(t+1)), \quad t=0..T \quad (5.15)$$

5.3. Решение и анализ демонстрационного примера

В качестве примера возьмем несколько тестов по машинному переводу (см. Приложение 3).

Посмотрим, какие слова попали в область значимых слов, а какие нет. В таблице 1 (см. Приложение 4) приведены все слова, входящие в тексты, указаны ранг частота и частота их вхождения, в таблицах 2, 3, 4 (см. Приложение 4) также приведены все слова для каждого текста и указана частота их вхождения (слова с частотой 1 опущены).

Слова с частотой 2 - 8 наиболее точно отражают смысл текстов. Слово с наибольшей частотой вхождения оказалось предлогом, а слова с меньшей - общими словами.

На рисунке 3 приведен график частота-ранг исследуемых текстов. Выделим зону значимых слов. Пусть это будут слова с рангом 2-8 и соответствующей частотой

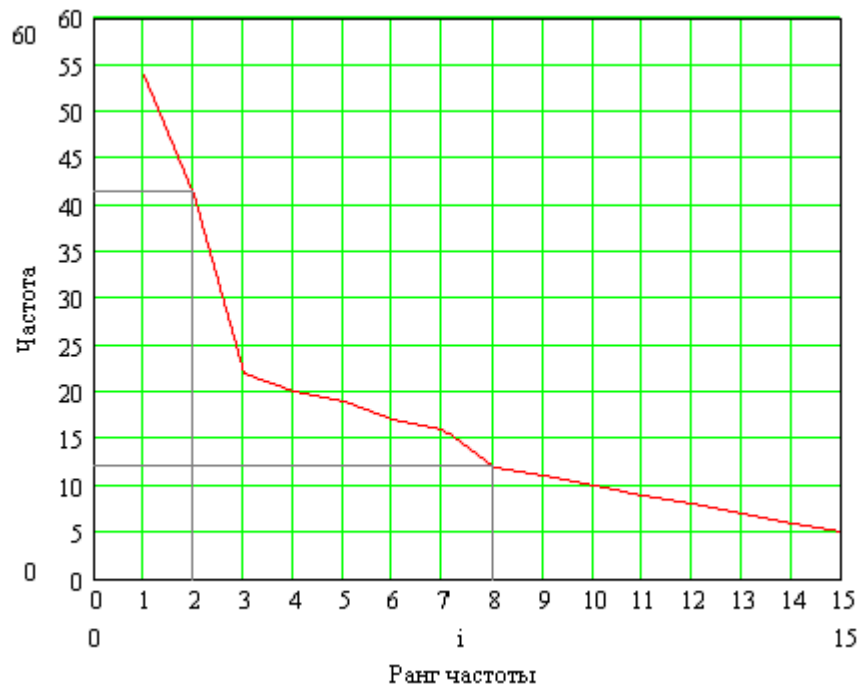


рис. 5.3. График частота-ранг исследуемых текстов

Проанализируем выделенную нами область значимых слов. Не все слова, которые попали в нее, отражают смысл текста.

В область попали и слова: и, на, что, это, которого. Эти слова являются "шумом", помехой, которая затрудняет правильный выбор. "Шум" можно уменьшить путем предварительного исключения из исследуемого текста некоторых слов. Для этого создается словарь ненужных слов -- стоп-слов (словарь называется стоп-лист). Для русского текста в стоп-лист могли бы быть включены все предлоги, частицы, личные местоимения и т. п.

Т.о, при использовании словаря ненужных слов область значимых слов будет состоять из слов с наибольшей частотой вхождения и какой-то средней частотой, определяемой лицом, принимающим решение либо определяться экспериментально.

Другой подход к уменьшению «Шума» - вычисление весов терминов.

Данный подход основывается на том, что, как правило, предлоги, частицы и т.п. встречаются в документах очень часто, в то время как слова, выражающие смысл текста, реже.

Т.о. термин, получивший нулевой или близкий к нулю вес, можно исключить из модели семантической сети.

Вычислим весовые коэффициенты терминов «в» и «программ»:

Инверсная частота термина «в» = $\log(3/3) = 0$

Вес термина «в» = $54 * \log(3/3) = 0$

Инверсная частота термина «программ» = $\log(3/2) = 0,1761$

Вес термина «в» = $19 * \log(3/2) = 3,3457$

То есть термин «в» можно исключить из модели семантической сети.

Так же вычислим весовой коэффициент слова «перевод»:

Инверсная частота термина «перевод» = $\log(3/3) = 0$

Вес термина «перевод» = $41 * \log(3/3) = 0$

Таким образом, получаем, что наиболее значимые, специфичные для данной темы слова, так же могут получить нулевой вес, поскольку могут встречаться в каждом из текстов данной тематики, так же как и предлоги. Однако, при большом числе исследуемых документов, маловероятно (хотя и не исключено), что значимые слова будут встречаться в каждом тексте базы данных. Разделение на значимые и незначимые термины (по их весовым коэффициентам) решает лицо, принимающее решение.

Составление словаря стоп-слов, представляется более надежным и простым, хотя и более трудоемким способом исключения «Шума». Кроме того, если учесть, что такие словари уже существуют в электронном виде, то исключение ненужных слов из модели семантической сети с помощью словаря является, по-видимому, более предпочтительным [3].

Программное обеспечение.

Для статистического анализа текстов (подсчета частот слов в тексте) использовалась бета-версия программы WordStat 1.2, распространяемая условно-бесплатно ("shareware").

Выводы

В главе была представлена статистическая модель порождения естественно-языкового текста. Рассмотрены вопросы применения модели для решения задачи классификации документов, в том числе оценка параметров модели на базе эталонных текстов.

Предложено рассматривать модель предметной области в форме сети, узлы которой представлены множеством часто встречающихся понятий текста - слов и устойчивых словосочетаний, из числа которых исключены общеупотребимые слова

Сделан анализ применения законов Зипфа для построения такой сети. Показано, как оценить вероятность того, что произвольный текст был порожден на основе заданной модели сети. На основе законов Зипфа приведен пример выделения наиболее значимых для заданной темы слов.

Заключение

Основные результаты работы:

1. Выполнен аналитический обзор существующих технологий построения систем машинного перевода. Рассмотрены различные классификации систем машинного перевода.
2. Разработана структура и основные принципы работы системы перевода с применением технологии памяти переводов.
3. Предложено объединить память переводов с системой словарей
4. Основываясь на анализе различных систем МП, разработана структура и основные принципы работы системы перевода с применением технологии машинного перевода.
5. Разработан алгоритм семантического анализа, основанный на использовании Русского общесемантического словаря (РОСС).
6. На основе описания РОСС разработан семантический словарь.
7. Разработана программа семантического анализа простых предложений на русском языке. В текущей версии программы семантическое представление предложения строится на основе анализа словарных статей семантического словаря.

Литература

1. Апресян Ю.Д. Избранные труды, Том 1. Лексическая семантика: 2-е изд., испр. и доп. - М.:Школа "Языки русской культуры" 1995.
2. Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л. и др. Лингвистическое обеспечение системы ЭТАП-3. //М., Наука, 1989. 295с.
3. Баскакова И.В. Классификация текстов по заданному набору тематик. Дни науки НГТУ-2003: Тезисы докладов студенческой конференции, Под ред. М.А.Кувшиновой. - Новосибирск: Изд-во НГТУ, 2003. - с.72.
4. Беляева Л.Н., Откупщикова М.И. Автоматический (машинный) перевод. – В сб.: Прикладное языкознание. СПб, 1996
5. Гак В.Г. Валентность. – Лингвистический энциклопедический словарь. М., 1990
6. Гак В.Г. Актант. – Лингвистический энциклопедический словарь. М., 1990
7. Глазунов А.Г. Концептно-ориентированная модель памяти переводов. – Центр Информационных Технологий,
<http://www.citforum.ru/programming/digest/cotm.shtml>.
8. Ермаков А.Е., Плешко В.В. Ассоциативная модель порождения текста в задаче классификации.// Информационные технологии. - 2000. - N 12.
9. Ермаков А.Е. Тематический анализ текста с выявлением сверхфразовой структуры // Информационные технологии. -2000.- N 11.
10. Крейдлин Л. М. Что такое UNL? //Компьютерра – 2001. - N 13.
11. Кулагина О. С. Машинный перевод: современное состояние. В сб.: Семиотика и информатика. Вып. 29. М., ВИНТИ, 1989.
12. Леонтьева Н.Н., Кудряшова И.М., Соколова Е.Г. Семантическая словарная статья в системе ФРАП // ПГЭПЛ. Ин-т русского языка АН СССР. Вып. 121. - М., 1979. 64 с.
13. Леонтьева Н.Н., Никогосов С.Л. Система ФРАП и проблема оценки качества автоматического перевода. - МГПИИЯ им. М. Тореца. Сборник научных трудов., Вып. 20., М.,1980.
14. Леонтьева Н.Н. Система французско-русского автоматического перевода (ФРАП): лингвистические решения, состав, реализация. - МГПИИЯ им. М. Тореца. Сборник научных трудов., Вып. 271., М.,1986.
15. Леонтьева Н.Н. "Политекст": информационный анализ политических текстов. //НТИ. Сер 2. – 1995.- N 4.- с 20-24.

16. Леонтьева Н.Н. Русский общесемантический словарь (РОСС): структура, наполнение. // НТИ. Сер. 2. - 1997. - N 12. - С.5-20.
17. Мельчук И.А. Опыт теории лингвистических моделей "Смысл <=> Текст". М.: Наука, 1974.
18. Попов Артем. Поиск в Интернете – внутри и снаружи. Эффективная методика поиска информации в сети Интернет. http://citforum.ints.net/pp/search_03.shtml
19. Raskin, V., Nirenburg S., Lexical Semantics of Adjectives, Recent Papers from the Mikrokosmos and Corelli Projects, Vol 2., New Mexico State University, 1996.
20. Сегалович Илья. Реализация словаря на основе разряженной хэш-таблицы.//Труды международного семинара Dialog' 95, Изд.Таруса 1995. <http://company.yandex.ru/articles/article5.html>
21. Слепов С.Н. Обзор технологий МП. <http://mt.slova.tk/review.htm>
22. Сокирко А.В. Реализация первичного семантического анализа в системе Диалинг. //Труды Международного семинара Диалог'2000 по компьютерной лингвистике и ее приложениям, Протвино, 1-5 июня 2000 года.
23. Сокирко А.В. Диссертация "Семантические словари в автоматической обработке текста (по материалам системы ДИАЛИНГ)", <http://www.aot.ru/technology.html>
24. Сонин О. М. МТ или ТМ. //Компьютерная неделя N 26-27.- М., 1999
25. Федоров А.В. Основы общей теории перевода. М., 1968
26. Филинов Е.Н. История машинного перевода.- Виртуальный компьютерный музей, <http://www.computer-museum.ru/histsoft/histmt.htm>

Приложения

Приложение 1

Список семантических характеристик

СХ	Комментарий	Примеры слов с таким СХ
АБСТР	Любое абстрактное существительное или прилагательное	модель план, тенденция, обстоятельство
АРТ	Артефакт. Все, что сделано человеком	машина, хлеб, памятник
ВЕЛИЧ	Прилагательные, образованные от параметрических существительных и от существительных, обозначающих какое-либо значение на параметрической шкале	высокий, мощный
ВЕЩВО	Любое название химического вещества или того, что можно как-либо дозировать, отмерять, продавать по весу или объему	аммиак, бензин
ВЛАСТЬ	Высшие государственные и военные должности и учреждения. Любые должности, связанные с непосредственным управлением людьми	генерал, президент
ВМЕСТЛ	Все, что предназначено для содержания чего-либо другого	мешок, гараж
ВОСПР	Все глаголы и существительные восприятия	слушать, видеть
ВРЕД	Все, к чему человек обычно относится как к нежелательному.	катастрофа, война.
ГЕОГР	Любой географический объект	остров, река.
ГОС	Любое название государства или тип государства	республика
ДВИЖ	Глаголы движения	идти, бежать

ДОЛЖ	Должность, профессия, социальный статус	повар, врач
Д-УСТР	Деталь устройства	валик
ИЗМ	Действия-изменения	реформировать
ИНТЕЛ	Все действия, непосредственно связанные с мыслительной деятельностью	изучать, решать.
ИНТРВЛ	Временной интервал	день, месяц.
ИНФ	Слова, обозначающие информацию	знание, новость
КОММУНИК	Глаголы речи	выражать, выступать.
НОСИНФ	Носители информации. Это можно прочесть, а потом сжечь.	книга, газета.
Н-ТРЕБ	Набор требований. (разновидность НОСИНФа)	закон, инструкция.
ОДЕЯТ	Область деятельности	физика, балет.
ОДУШ	Семантически одушевленный объект	мама, брат.
ОРГ	Любая организация	школа.
ОТН	Слово-отношение. Как правило, валентная структура такого слова состоит из П-АКТ и В-АКТ	включать, быть
ОЦЕНКА	Эти слова обычно имеют морфологические степени сравнения и предполагают некоторого субъекта-оценщика, который проявляется в контекстах: «красивый для меня», «я нахожу его красивым».	хороший, плохой
ОТСУТ	Отсутствие какого-либо признака	избавляться
ПРЕДМ	Любой предмет (объект, который меньше по размерам среднего человека). Часто является АРТ. В отличие от УСТР устроен просто.	марка, бинокль
ПРОТЯЖ	Протяженные географические объекты	дорога, граница.

СИТУАТ	Слово-ситуация, для СХ - главного слова: СИТУАТ синонимично категории ЭТК.СИТ	бегать
СОБИР	Все, что обозначает множество однотипных объектов	библиотека, молодежь
СОЦ	Любые ситуации, выходящие за пределы одной семьи	митинг, олимпиада
СУЩЕСТ	Характеристика, означающее существование объекта в мире	
УСТР	Любое устройство	компьютер, лифт
ФИН	Все, что связано с финансами	банк, деньги
ХОР	Все, что оценивается как положительное	мужество, помощь
ЭМОЦ	Обычно прилагательные, которые выражают эмоции	могучий, несчастный
ЯВЛЕН	Ситуация, для которой трудно найти причину	смерч, терроризм

Приложение 2

Список валентностей

Название	Примеры	Структура
АВТОР	Роман Толстого	АВТОР(Толстой, роман)
АГЕНТ	Мы сократили отставание	АГЕНТ(мы, сократить)
АДР	Я отдал стул отцу	АДР(отец, отдавать)
АКТ	Болезнь сблизила их	АКТ (их, сблизить)
В-АКТ	Эта потеря сблизила брата с сестрой	В-АКТ(сестра, сблизить)
В-НАПР	Указатель на Монино	В-НАПР(Монино, указатель)
ВРЕМЯ	Это произошло вчера	ВРЕМЯ(Вчера, Произойти)
ЗНАЧ	Высота дома – 20 метров	ЗНАЧ(20 метров, высота)
ИДЕНТ	Дом N 20	ИДЕНТ (N 20, дом)
ИНСТР	Резать ножом	ИНСТР(нож, резать)
ИСХ-Т	Яблоки из Молдавии	ИСХ-Т(Молдавия, яблоки)
К-АГЕНТ	Купил у старьевщика	К-АГЕНТ(старьевщик, купить)
КОЛИЧ	Два яблока	КОЛИЧ(два, яблоко)
КОН-Т	Уехать в Москву	КОН-Т(Москва, уехать)
ЛОК	Жить в глуши	ЛОК(глушь, жить)
МАСШТ	Банк России	МАСШТ(Россия, банк)
МАТЕР	Сумка из кожи	МАТЕР(кожа, сумка)
НАЗН	Книга для детей	НАЗН(дети, книга)
ОБ	Уничтожить мост	ОБ(мост, уничтожить)

ОБ.РЕЗЛТ	Строить дом	ОБ.РЕЗЛТ(дом, строить)
ОГРН	Выделять по возрасту	ОГРН(возраст, выделение)
ОЦЕНКА	Хорошо относиться	ОЦЕНКА(хорошо, относиться)
П-АКТ	Эта потеря сблизила брата с сестрой	П-АКТ (брат, сблизить)
ПАРАМ	Высота дома	ПАРАМ(высота, дом)
ПАЦИЕН	Арест преступника	ПАЦИЕН(преступник, арест)
ПОСРЕД	Закончить доклад анекдотом	ПОСРЕД(анекдот, закончить)
ПРИЗН	Красивый шар	ПРИЗН(красивый, шар)
ПРИНАДЛ	Дом отца	ПРИНАДЛ(дом, отец)
ПРИЧ	Деревья повалены ураганом	ПРИЧ(ураган, повалить)
РЕЗЛТ	Испечь пирог	РЕЗЛТ(пирог, испечь)
СОДЕРЖ	Рассказать о весне	СОДЕРЖ(весна, рассказать)
СПОСОБ	Идти босиком	СПОСОБ(босиком, идти)
СРЕДСТВО	Красить белилами	СРЕДСТВО(белило, красить)
СТЕПЕНЬ	Весьма преуспеть	СТЕПЕНЬ(весьма, преуспеть)
СУБ	Любовь отца	СУБ(отец, любовь)
ТЕМА	Говорить о Москве	ТЕМА(Москва, говорить)
ЦЕЛЬ	Забастовка в целях повышения зарплаты	ЦЕЛЬ (повышение, забастовка)
ЧАСТЬ	Ножка стула	ЧАСТЬ(ножка, стул)

Приложение 3

Тексты по машинному переводу

Современные технологии

Программы машинного перевода пережили недавно возрождение, во многом связанное с развитием Интернета и с ростом его доступности для всё большего числа людей. Существует множество программ для перевода и для переводчиков. Интересно, что полезность таких программ определяется не только качеством перевода, который делает машина, но и открытостью программы, лёгкостью её интеграции с другими продуктами обработки документов (прежде всего с текстовыми редакторами и Интернет-браузерами).

В частности, существуют браузеры со встроенной функцией перевода. Выглядят они как обычный навигатор, окно которого разделено пополам: в одной половине отображается оригинал, а в другой – перевод. Причём процесс перевода занимает считанные секунды (так как объём веб-страниц часто невелик), а переведённый документ сохраняет разметку оригинала (расположение текста и рисунков на странице, вид шрифтов, гиперссылки и т.п.) Есть также онлайн-сервисы, делающие то же самое, но при этом не требующие установки программы на вашем компьютере.

Сравнительно недавно появился ещё один вид программ для перевода. Вернее, для людей-переводчиков. Они основаны на технологии Translation Memory, ТМ (в противоположность МТ, машинному переводу). Идея заключается в хранении базы данных переводов, сделанных профессиональным переводчиком для того, чтобы в процессе перевода предлагать человеку уже готовый перевод фразы или куска текста,

если он уже был однажды переведён. Причём совпадение фразы не обязательно должно быть буквальным, а может определяться критериями "похожести", заложенными в программу, с возможностью их настройки пользователем. ТМ-программы очень полезны в ситуациях, когда необходимо сделать перевод обновлённой версии документа, переведённого ранее. Такая необходимость возникает при поддержке мультязычных сайтов. Программа быстро обнаружит в документе места, подвергшиеся изменениям со времени предыдущей версии документа, и человеку останется перевести только эти изменившиеся части. ТМ-программы значительно повышают эффективность работы переводчика, избавляя его от рутинной, повторяющейся работы. Во многих фирмах, занимающихся переводом, владение одной из таких программ является существенным критерием при приёме на работу.

Существуют программы, объединяющие технологии МТ и ТМ.

Категории машинного перевода

В 1990 году Ларри Чаилдсом была предложена классификация систем машинного перевода:

- FAMT (Fully-automated machine translation) — полностью автоматизированный машинный перевод;
- HAMT (Human-assisted machine translation) — машинный перевод, сделанный при участии человека;
- MANT (Machine-assisted human translation) — перевод, осуществляемый человеком с использованием компьютера.

Идеальным средством для технического перевода мог бы оказаться переводчик, построенный только по схеме FAMT. Однако в ближайшие годы чисто машинный перевод едва ли найдет серьезное практическое применение в силу сложности, многообразия и недостаточной "формализуемости" естественных языков.

Программы, построенные по схеме HAMT, разработчики называют МТ-программы (от Machine translation - машинный перевод). Реально автоматизированный (с участием человека) машинный перевод возможен только в условиях искусственно ограниченного, как по словарному запасу, так и по грамматике, языка. В качестве реального успешного проекта МТ-программы всегда называют немецкую систему Meteo, выполняющую перевод метеопрогнозов с французского языка на английский и обратно.

MANT является самым трудоемким (с точки зрения переводчика), однако он кажется наиболее надежным, поскольку кто, если не человек, способен адекватно передать смысл, заключенный в тексте? Тем не менее, более внимательный взгляд на проблему позволяет обнаружить, что человеческий перевод по-настоящему бесценен только в художественной литературе и публицистике, где важными факторами являются разнообразие и творческий подход. В то же время, научные и технические тексты требуют строгих формулировок и точного следования терминологии, что временами представляет для человека проблему.

Программы, построенные по схеме MANT, разработчики называют ТМ-программы (от translation memory - память перевода). Эту категорию программ применяют профессиональные переводчики, осознавшие выигрыш от автоматизации их работы с помощью компьютеров. Основу ТМ-программ составляют специализированные словари, соответствующие тематике переводимого текста. При переводе используются конструкции и значения слов и устойчивых словосочетаний, выбранные профессиональным переводчиком и занесенные в словари системы, а полученный текст подвергается интенсивному редактированию. Словари и уже переведенные фрагменты текстов, запоминаемые в ТМ-системе, могут быть повторно использованы в больших

коллективных проектах, ими можно обмениваться. Поэтому ТМ-системы представляют собой важное средство автоматизации труда профессиональных переводчиков.

Переведутся ли переводчики?

Определение МП

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. Современный машинный, или автоматический перевод осуществляется с помощью человека: пред-редактора, который тем или иным образом предварительно обрабатывает подлежащий переводу текст, интер-редактора, который участвует в процессе перевода, или пост-редактора, который исправляет ошибки и недочеты в переведенном машиной тексте.

Немного истории

Автоматический ("машинный") перевод текстов исторически был одной из первых задач, решение которых люди попытались переложить на вычислительные устройства. По-видимому, первым, кто попытался получить правительственные субсидии на развитие вычислительной техники, был выдающийся математик XIX века Чарльз Бэббидж. В числе благ, которые он сулил британскому правительству в случае поддержки его проекта вычислительной машины, было обещание, что когда-нибудь эта машина сможет автоматически переводить разговорную речь.

Другие изобретатели тоже пытались создать механические переводящие устройства еще до наступления компьютерной эры. Например, Петр Троянский в середине 1930-х годов получил в СССР патент, предложив не только автоматический двуязычный словарь, но и схему кодирования межъязыковых грамматических ролей, основанную на языке эсперанто. Тем не менее, сейчас принято считать, что основные принципы современного машинного перевода были изложены только в 1947 году в письме директора естественнонаучного отделения Рокфеллеровского фонда Уоррена Уивера к Норберту Винеру.

За этим письмом последовала активная дискуссия среди специалистов, а уже через пять лет был переведен знаменитый Джорджтаунский эксперимент, имевший грандиозный успех. В ходе него был продемонстрирован электронный словарь, содержащий всего 250 слов и шесть грамматических правил. Это обеспечивало перевод полусотни заранее отобранных предложений.

После этого эксперимента возможности компьютерного перевода рассматривались в самом радужном свете, а будущее переводчиков-профессионалов, наоборот, представлялось очень и очень проблематичным. Однако уже в 1966 году консультативный комитет по автоматической обработке языка при Национальной академии наук США (*ALPAC*) представил крайне пессимистический отчет о перспективах машинного перевода, после чего почти все работы в этой области были свернуты и практически заморожены до самого конца 1970-х годов (причем не только в США, но и в СССР, и в большинстве стран Европы). Только падение "железного занавеса", развитие международной коммерции и Интернета дали новый мощный толчок (подкрепленный финансовыми вливаниями) для исследований в этой сфере.

С середины 90-х годов перевод веб-страниц "на лету" постепенно становится одной из приоритетных задач всех систем машинного перевода. При этом, конечно, никто всерьез не рассматривает "чисто машинный перевод" как окончательный. Основные работы сейчас ведутся в сферах, которые принято обозначать аббревиатурами МАНТ (*Machine-Aided Human Translation*, человеческий перевод с привлечением машин) и НАМТ (*Human-Aided Machine Translation*, машинный перевод с участием человека).

"А в чем, собственно, проблема?"

Первый источник проблем машинного перевода - это многозначность слов в любом естественном языке (профессиональные лингвисты стараются различать полисемию и омонимию, но, с точки зрения переводчиков, и то, и другое приводит к одинаковым трудностям), а также существование устойчивых словосочетаний и фразеологических оборотов. Причем, эти явления существуют как в языке, с которого делается перевод, так и в том языке, на который переводится. Вот один из примеров, увиденный автором буквально на днях. Слоган фирмы Western Union - "The fastest way to transfer money world wide" переведен на русский язык так: "*Western Union* - самый быстрый способ перевести деньги по всему миру". Что называется, "попали"... Во-первых, оборот "перевести деньги" имеет еще и значение "протратиться" (сравните: "деньги у него совсем перевелись"), а во-вторых, очень давит схожесть конструкции "перевести деньги по миру" с русской идиомой "пустить по миру". Более явную антирекламу придумать очень сложно... (Зато запоминается. Может быть, в этом и состояла задумка переводчика?!)

Второй источник погрешностей при переводе - требования языка к соблюдению определенного порядка слов в предложениях, то есть к способу объединения отдельных слов в связный текст.

Принято считать, что в русском языке порядок слов свободный: вы можете как угодно переставить слова в фразе "я выпил молоко" - и собеседник вас поймет однозначно. Исключения бывают, но сравнительно нечасто ("мать любит дочь" не то же самое, что "дочь любит мать"). В то же время, в английском, да и в большинстве европейских (германских и романских) языков, соблюдение вами правильного порядка слов жизненно необходимо для того, чтобы ваш собеседник смог понять, что же вы ему пытались сообщить.

И, наконец, третий источник лингвистических затруднений - невозможность формально описать лингвистические закономерности. Например, школярские представления о том, что в русском языке существует всего 36 категорий имени существительного (три рода, три склонения, две категории одушевленности, имена собственные/нарицательные), увы, совершенно не подтверждаются живым языком. Слова "глаз", "луч", "матрац", "стул" и "стол" любой школьник отнесет к нарицательным существительным мужского рода, второго склонения, неодушевленным. Однако в именительном падеже множественного числа будут "глазА", "лучИ", "матрацЫ", "стульЯ", а в родительном падеже множественного числа разнообразие вариантов еще больше: "глаз", "лучЕЙ", "матрацЕВ", "стульЕВ", "столОВ".

Поэтому не стоит ожидать от машинного перевода больше, чем он в принципе может дать. Конечно, рано или поздно машинные переводчики достигнут того уровня, когда улучшить и поправить их перевод сможет только настоящий профессионал. Но сейчас говорить об этом все еще преждевременно: грубые ошибки видны невооруженным глазом.

Приложение 4

Статистический анализ текстов

Ранг	Частота	Слова
1	54	В
2	41	И
2	41	ПЕРЕВОД
3	22	МАШИННОГО
4	20	ПЕРЕВОДИТСЯ
5	19	ПРОГРАММ
6	17	С
7	16	НА
7	16	ЯЗЫК
8	12	НЕ
8	12	ПО
8	12	СЛОВ
8	12	ТЕКСТ
8	12	ЧТО
9	11	А
9	11	ТАК
9	11	ТОЛЧОК
10	10	ДЛЯ
10	10	КОТОРОГО
11	9	БЫЛ
11	9	ЧЕЛОВЕК
11	9	ЭТО
12	8	TRANSLATION
12	8	ПРИ
13	7	ГОДОВ
13	7	НО
14	6	MACHINE
14	6	КАК
14	6	РАБОТУ
14	6	САМОГО
14	6	СИСТЕМ
14	6	ТО
14	6	ЧИСЛА
15	5	АВТОМАТИЧЕСКИ
15	5	БОЛЕЕ
15	5	ВСЕГДА
15	5	ДОКУМЕНТ
15	5	ДРУГИЕ
15	5	ЖЕ

Таблица 1. Статистический анализ текстов.

15	5	ИЛИ
15	5	К
15	5	МОЖЕТ
15	5	ОДНАЖДЫ
15	5	ОДНОЙ
15	5	ОСНОВАННУЮ
15	5	ОТ
15	5	ОЧЕНЬ
15	5	ПРОФЕССИОНАЛЬНЫЕ
15	5	СУЩЕСТВУЕТ
15	5	УЖЕ
16	4	HUMAN
16	4	ТМ
16	4	ВО
16	4	ВРЕМЕНАМИ
16	4	ГЛАЗ
16	4	ДЕНЬГИ
16	4	ЕЩЕ
16	4	ИЗ
16	4	КАТЕГОРИИ
16	4	ЛЮБИТ
16	4	ОН
16	4	ПЕРЕВЕСТИ
16	4	ПРИЧЕМ
16	4	РЕДАКТОРА
16	4	РУССКИЙ
16	4	СТРАН
16	4	СХЕМ
16	4	ТМ
17	3	НАМТ
17	3	МАНТ
17	3	БЫТЬ
17	3	ВЕРНЕЕ
17	3	ВТОРОГО
17	3	ВЫЧИСЛИТЕЛЬНОЙ
17	3	ЕГО
17	3	ИМЕЕТ
17	3	ИНТРЕНЕТ
17	3	ИСТОЧНИК
17	3	ИХ

17	3	КОГДА
17	3	КОМПЬЮТЕРА
17	3	КОМПЬЮТЕРНОГО
17	3	ЛУЧ
17	3	ЛЮДЕЙ
17	3	МАТРАЦ
17	3	МАШИН
17	3	МИРУ
17	3	НАЗЫВАЮТ
17	3	ПАДЕЖЕ
17	3	ПЕРЕВЕДЕН
17	3	ПОРЯДКА
17	3	ПОСТРОЕННЫЕ
17	3	ПРЕДЛОЖЕНА
17	3	ПРИНЯТО

17	3	ПРОЕКТА
17	3	ПРОЦЕСС
17	3	РАЗВИТИЕ
17	3	СЕЙЧАС
17	3	СМОГ
17	3	СОВРЕМЕННОГО
17	3	СПОСОБ
17	3	СТОИТ
17	3	ТЕМ
17	3	ТЕХНОЛОГИИ
17	3	ТОГО
17	3	УЧАСТИЕМ
17	3	ЭТИ

Таблица 2. Статистический анализ текста «Современные технологии»

Частота	Слова
13	ПЕРЕВОД
12	ПРОГРАММ
9	В
8	И
6	ДЛЯ
5	ДОКУМЕНТ
5	С
5	ТАК
4	ТМ
4	ПРИ
4	НА
4	ПЕРЕВОДЧИКА
3	А
3	ВЕРНЕЕ
3	НЕ
3	ПРИ
3	ПРИЕМЕ
3	РАБОТУ
3	СУЩЕСТВУЕТ
3	ТЕХНОЛОГИИ
2	МТ
2	БРАУЗЕРАМИ
2	ВИД
2	ВО

2	ДРУГИМИ
2	ЕГО
2	ИНТЕРНЕТ
2	КАК
2	КОТОРОГО
2	КРИТЕРИЕМ
2	ЛЮДЕЙ
2	МАШИННОГО
2	МНОГИХ
2	НЕДАВНО
2	НО
2	ОДНОЙ
2	ОНИ
2	ОРИГИНАЛ
2	ПЕРЕВЕДЕННОГО
2	ПРОЦЕСС
2	СО
2	СТРАНИЦ
2	ТЕКСТА
2	ТОЛЬКО
2	УЖЕ
2	ФРАЗЫ
2	ЧЕЛОВЕКУ
2	ЧТО

Таблица 3. Статистический анализ текста «Категории машинного перевода».

Частота	Слова
13	ПЕРЕВОД
11	И
11	В
7	МАШИННОГО
7	ПРОГРАММ
5	TRANSLATION
5	С
5	ПРОГРАММЫ
5	TRANSLATION
5	ПЕРЕВОДЧИК
5	ПОСТРОЕННЫЕ
5	С
5	СИСТЕМ
5	ТЕКСТ
5	ЧЕЛОВЕК
4	MACHINE
4	TM
3	МАНТ
3	НАЗЫВАЮТ
3	ОТ
3	ПОСТРОЕННЫЕ
3	ПРОФЕССИОНАЛЬНЫЕ
3	СЛОВАРИ
3	СХЕМЕ
3	ТОЛЬКО
3	ЯЗЫКА
2	ASSISTED

2	FAMT
2	НАМТ
2	HUMAN
2	АВТОМАТИЗАЦИИ
2	АВТОМАТИЗИРОВАННЫЙ
2	ВАЖНОЕ
2	ГОДУ
2	ДЛЯ
2	ИСПОЛЬЗОВАНИЕМ
2	КАТЕГОРИИ
2	КОМПЬЮТЕРА
2	МОГ
2	МТ
2	НА
2	НЕ
2	ОДНАКО
2	ПРЕДСТАВЛЯЕТ
2	ПРИ
2	ПРОБЛЕМУ
2	ПРОЕКТА
2	РАЗРАБОТЧИКИ
2	РЕАЛЬНО
2	СРЕДСТВО
2	ТЕХНИЧЕСКИЕ
2	УЧАСТИЕМ
2	ЧТО

Таблица 4. Статистический анализ текста «Переведутся ли переводчики?»

Частота	Слова
34	В
22	И
18	ПЕРЕВОД
13	МАШИННОГО
13	ЯЗЫК
10	НА
10	СЛОВ
8	КОТОРОГО
8	ЧТО
8	ЭТО
7	А
7	БЫЛ
7	НЕ
7	ПО
7	С
6	ТЕКСТ
6	ТОЛЧОК
5	АВТОМАТИЧЕСКИ
5	ВСЕ
5	ГОДОВ

5	К
5	НО
5	ПЕРЕВОДЧИКА
5	ТАК
4	ГЛАЗ
4	ДЕНЬГИ
4	ЕЩЕ
4	ИЛИ
4	ЛЮБИТ
4	ОЧЕНЬ
4	ПЕРВЫЙ
4	РУССКИЙ
4	САМОГО
4	ТО
4	ЧИСЛА
3	БОЛЕЕ
3	ВТОРОГО
3	ВЫЧИСЛИТЕЛЬНОЙ
3	ДРУГИЕ
3	ЖЕ
3	ИЗ
3	ИМЕЕТ

3	ИСТОЧНИК
3	КАК
3	ЛУЧ
3	МАТРАЦ
3	МИРУ
3	МОЖЕТ
3	ОДНОЙ
3	ПАДЕЖЕ
3	ПЕРЕВЕСТИ
3	ПЕРЕВОДИТСЯ
3	ПОРЯДКА
3	ПРИ
3	ПРИНЯТО
3	РЕДАКТОРА
3	СЕЙЧАС
3	СМОГ
3	СТОИТ
2	AIDED
2	HUMAN
2	MACHINE
2	TRANSLATION
2	UNION
2	WESTEN
2	БОЛЬШИНСТВЕ
2	БУДУТ
2	ВО
2	ВЫ
2	ГРАММАТИЧЕСКИХ
2	ДЕЙСТВИЕ
2	ДЛЯ
2	ДО
2	ДОЧЬ
2	ЕСТЕСТВЕННОМ
2	ЗАДАЧ
2	КАТЕГОРИИ
2	КОГДА
2	КОМПЬЮТЕРНОГО
2	КОНЕЧНО
2	ЛЕТ
2	ЛИНГВИСТИЧЕСКИЕ
2	МАТЬ
2	МАШИН
2	МНОЖЕСТВЕННОГО
2	НАПРИМЕР
2	НАРИЦАТЕЛЬНЫЕ
2	НЕГО
2	О
2	ОБОРОТ
2	ОДНАКО
2	ОН
2	ОСНОВНЫЕ
2	ОШИБКИ
2	ПЕРЕВЕДЕН
2	ПИСЬМЕ

2	ПОЛУЧИ
2	ПОПЫТАЛИСЬ
2	ПОСЛЕ
2	ПРИДЛОЖЕНИЙ
2	ПРИНЦИПЕ
2	ПРИЧЕМ
2	ПРОБЛЕМ
2	ПРОФЕССИОНАЛ
2	ПЫТАЛИСЬ
2	РАБОТЫ
2	РАЗВИТИЕ
2	РОДА
2	СЕРЕДИНЕ
2	СКЛОНЕНИЯ
2	СЛОГАН
2	СОБЕСЕДНИК
2	СОБЛЮДЕНИЕ
2	СОБСТВЕННО
2	СОВРЕМЕННОГО
2	СПОСОБ
2	СССР
2	СТРАН
2	СТУЛЬЕВ
2	СУЩЕСТВИТЕЛЬНОГО
2	СУЩЕСТВУЕТ
2	СФЕРАХ
2	СЧИТАТЬ
2	США
2	ТЕМ
2	ТОГО
2	ТОМ
2	ТРИ
2	УЖЕ
2	УСТРОЙСТВА
2	ЧЕЛОВЕКА
2	ЧЕМ
2	ЭКСПЕРИМЕНТ
2	ЭТИ

Приложение 5

Исходные тексты программ