



Системы машинного перевода и
смысловой обработки текстов,
информационно-аналитические и
поисковые системы

Методы искусственного интеллекта (концептуальные решения)

MetaFraz 2020 г.

Методы искусственного интеллекта

В НИР «ВЕЛЕС» были определены базовые технологии обработки неструктурированной текстовой информации, являющиеся основой для развития искусственного интеллекта в этой области:

1. Машинные грамматики естественных языков
2. Средства смысловой обработки текстов (семантико-синтаксический, дистрибутивно-статистический и концептуальный анализ текстов)
3. Автоматизированное построение терминологических и фактологических онтологий

Концепция фразеологического концептуального анализа

- В качестве базовой теоретической концепции использовалась **концепция фразеологического концептуального анализа текстов** констатирующая, что смысловое содержание текстов выражается с помощью единиц смысла, входящих в их состав и их смысловых отношений.
- Наиболее **устойчивой единицей смысла** является **понятие**, определяемое как социально значимый мыслительный образ, за которым в языке закреплено его наименование в виде отдельного слова или, значительно чаще, в виде устойчивого фразеологического словосочетания.

Концепция фразеологического концептуального анализа

- **Понятия** занимают центральное место в языке и речи и являются теми базовыми строительными блоками, на основе которых формируются смысловые единицы более высоких уровней.
- **Смысл понятия** проявляется в полной мере только через всю систему его отношений со всеми другими понятиями языка.
- Второй по значимости единицей смысла является **предложение**. Из предложений формируются различного рода **сверхфразовые единства**, которые представляются в виде последовательностей связного текста.
- Смысловое содержание предложения выражается через его **предикатно-актантную структуру**.

Концепция фразеологического концептуального анализа

- **В связном тексте** предложения выступают в тесной смысловой связи. В основе этой связи лежат мыслительные образы тех конкретных или абстрактных объектов (ситуаций, явлений), которые человек имеет в виду, когда порождает текст.
- **Локальная связность** обеспечивает раскрытие смысла понятия на основе его контекста. Под смысловой связанностью текста или его фрагмента будем понимать совокупность наименований понятий, расположенных в тексте в определённом порядке и отражающих основное смысловое содержание текста или его фрагмента.

Концепция фразеологического концептуального анализа

- **Локальная смысловая схожесть наименований понятий** текста определяется как сходство контекстного окружения идентичных наименований понятий в двух текстах или их фрагментах.
- **Глобальная смысловая схожесть текстов** или их фрагментов определяется как сходство состава идентичных наименований понятий и порядка их следования в текстах или их фрагментах. Каждое понятие этого фрагмента также должно удовлетворять условию локальной смысловой схожести

Основная единица смысла – терминологическое или фразеологическое понятие

Понятия в текстах выражаются **словами**, но чаще **словосочетаниями**.

По данным международного терминологического центра **ИНФОТЕРМ** в развитых естественных языках (англ, русск, нем., и др):

- более одного миллиона лексем (7 – 15 млн. словоформ)
- более 50 млн. терминов (в различных формах представления)
- более 100 млн. - номенклатура изделий и товаров

При автоматическом распознавании понятий в текстах должна учитываться **вариативность форм их представления**.

Машинные грамматики

1. Машинные грамматики (МГ) относятся к базовым процедурам обработки текстовой информации и являются «входными воротами» любой интеллектуальной системы.
2. МГ предназначены для автоматического установления морфологической структуры слов и определения полного набора их грамматических характеристик (ГХ).
3. Основное требование, предъявляемое к МГ – правильность назначения ГХ всей совокупности слов русского языка (7-15 млн.) и высокое быстродействие.

Машинные грамматики (традиционные)

1. Традиционно процедуры машинной грамматики строились на подходах, базирующихся на использовании **словарей слов или их основ**, для которых заранее был назначен **набор грамматических характеристик**.
2. Такой подход при всех его преимуществах по обеспечению возможности назначения правильных грамматических характеристик слов требовал **больших первоначальных трудозатрат** при создании словарей большого объема, а также **значительных машинных ресурсов** при реализации этого подхода - большого числа итераций при поиске основы анализируемого слова в словаре основ и проверки на совместимость найденной основы со списком возможных окончаний.

Машинные грамматики (проблемы)

1. При традиционном подходе неизбежно возникает **проблема обработки «новых» слов** или иными словами - проблема определения грамматических характеристик тех слов, которые не вошли в состав словаря (200 тыс. слов словаря из 7-15 млн. слов всей совокупности слов русского языка).
2. Обычно обработку «новых слов» связывали с **анализом суффиксов (сочетаний суффиксов и окончаний**. Между тем многие формы слов не имеют суффиксов и окончаний, и анализ таких слов с помощью такого подхода не привел к успеху.

Машинные грамматики (История)

В середине 60-ых годов прошлого века сотрудник 27 ЦНИИ - проф. полковник Г.Г. Белоногов проводил исследования по анализу лексического состава текстов военного назначения. В ходе этих исследований он установил:

- В русском языке есть **четкая взаимосвязь между грамматическими характеристиками слов русского языка и их буквенным составом.**
- По характеру изменения грамматических окончаний (флексий) и по своей синтаксической функции русские слова могут быть разбиты на ряд классов, которые получили название **флексивных.**
- Всю совокупность слов русского языка можно разделить на две группы – слова с **регулярной системой словоизменения и словообразования (90-95%)** и **нерегулярной (5-10%)** – все

Машинные грамматики (решение проблемы)

1. Революционное решение проблемы было предложено в 1990 г. в рамках научной школы проф. Белоногова
2. Было предложено «старые» словоформы русского языка – служебные слова, короткие слова (менее 5 букв), часть имен и фамилий с аномальной системой словоизменения и словообразования включить в словарь коротких и служебных слов (КСС)- (объем 80-100 слов)
3. Все остальные слова, подчиняющихся регулярным законам словоизменения и словообразования рассматривать как «новые» слова и вычислять их ГХ на основе метода лингвистической аналогии (скажи - кто твои друзья и я скажу - кто ты)

Машинные грамматики (решение проблемы)

1. Это решение базируется на уникальной классификации слов русского языка, разработанной в середине 60 годов прошлого века проф. Г.Г. Белоноговым и названной таблицей флективных классов (ФК) русского языка.
2. Эту компактную таблицу по своей значимости можно назвать «таблицей Менделеева» для русского языка.
3. Вторым открытием, которое позволило автоматизировать процессы создания декларативных средств является установление корреляционной связи между конечным буквосочетанием слова (КБС), его морфологической структурой (МС) и набором ГХ.
4. В процессе исследований была также установлена жесткая взаимозависимость между ФК слова, его грамматического окончания (Ок) и набора ГХ. Это был прорыв в представление формальных конструкций предложения (ФКП).

Машинные грамматики (Выводы)

Основополагающими принципами и методами создания машинных грамматик языков ряда стран Западной и некоторых языков стран Юго-восточной Азии являются следующие:

- **Статистические методы анализа** разноязычных репрезентативных корпусов текстов, позволяющие выявить лексический состав национальных языков и соотнести его с параметром покрытия этой лексикой текстов на этом языке.
- **Принцип лингвистической аналогии** позволяет выявлять и реализовать трансформационные закономерности словоизменения и словообразования, многократно сократить объемы словарей и грамматических таблиц, а также успешно решить задачи, которые не поддающиеся решению алгоритмическими методами.

Машинные грамматики (Выводы)

- Созданный **комплекс декларативных средств машинной грамматики** полностью покрывает все возможные реализации по автоматическому определению структуры слов, установлению их грамматических характеристик, а также средств установления смыслового тождества между членами одной словоизменительной парадигмы.
- Широкое использование **средств автоматизации при формировании декларативных средств** позволили многократно сократить трудозатраты и существенно повысить качество созданных словарных ресурсов.
- Разработанные программные средства базируются на **уникальных алгоритмах, обеспечивающих их быстроедействие и высокое качество** обработки текстовой информации.

Семантико-синтаксический анализ текстов

- В процессе семантико-синтаксического анализа (ССА) выполняется построение семантико-синтаксической модели текста.
- ССА - это лингвистическая процедура, обеспечивающая определения в тексте различных синтаксических конструкций и установления синтаксических связей между ними.
- Исходными данными для ССА являются результаты графематического и морфологического анализа текстов.
- Результатами работы ССА является формальная синтаксическая модель текста.

Синтаксическая модель на основе обобщенных синтагм

- В рамках концепции фразеологического анализа текстов было установлено жесткое соответствие между формой представления слов и их грамматической информацией, которое позволило создать на этой основе новые классы – *классы слов, имеющие одинаковые наборы грамматических признаков, соответствующие их формам представления в сходных контекстных окружениях.*
- Идея создания новых классов слов, ориентированных на схожесть грамматических признаков слов и схожесть их синтаксических функций в предложении, была ранее предложена для разрешения грамматической омонимии английских слов.

Синтаксическая модель на основе обобщенных синтагм

- При разработке этой синтаксической модели текстов мы исходили из следующей гипотезы: **одинаковым последовательностям обобщенных символов классов слов (обобщенным синтагмам) должны соответствовать одинаковые синтаксические структуры.**
- Представление предложения в виде последовательности индексов обобщенных синтагм по сути представляет **синтаксическую структуру предложения, элементами которой являются контактно расположенные объекты, обладающие грамматическими свойствами конкретных слов-эталонов. Конкретным элементам структуры могут соответствовать различные слова, грамматические признаки которых идентичны.**

Представление синтаксической структуры предложения в виде индексов обобщенных синтагм

Трубопроводы высокого давления резервной котельной расположены на значительном расстоянии от систем безопасности блока -> S= трубопровод; P= расположен на; O= расстояние.

Aq Цf ЙЦ Фh Фh жq 7A Фi ЙS яA uA tS FB -> S=Aq; P= жq 7A; O= яA.

- Aq - Сущ., муж. р. , неодуш., ФК=001 1)мн. ч., им. п. 2) мн. ч., вин. п.
- Цf - Полн. прил., ФК=106 1)муж. р. , ед. ч, род п. 2) муж. р. , ед. ч, вин. п. 3) ср. , ед. ч, род. п.
- ЙЦ - Сущ., ср. р. ,неодуш., ФК=073 1)ед. ч., род. п. 2) мн. ч., им. п. 2) мн. ч., вин. п.
- Фh - Полн. прил., ФК=103 жен. р. , ед. ч, род п. 2) жен. р. , ед. ч, дат. п. 3) жен. , ед. ч, тв. п. 4) жен. , ед. ч, пред. п.
- Фh - Полн. прил., ФК=103 жен. р. , ед. ч, род п. 2) жен. р. , ед. ч, дат. п. 3) жен. , ед. ч, тв. п. 4) жен. , ед. ч, пред. п.
- жq - Кратк. прич., ФК=126, мн. ч.
- 7A - Предлог, ФК=164, мод. упр.- вин. п., пред. п.

Семантико-синтаксический анализ текстов (процесс)

Семантико-синтаксический анализ текстов выполняет следующие задачи по созданию семантико-синтаксической модели текста:

- Реализует модель дерева зависимостей
- Реализует модель членов предложения
- Реализует модель непосредственно составляющих
- Устанавливает границы именных и глагольных словосочетаний в предложении
- Определяет главные и зависимые слова в словосочетании и устанавливается тип связи между ними
- Определяет синтаксическую роль слов в предложении

Семантико-синтаксический анализ текстов (процесс)

- Строит «скелет» предложения – определяет главные слова словосочетаний и устанавливает их взаимосвязи в предложении
- Выявляет предикатно-актантную структуру предложения
- Строит универсальное представление структуры предложения в виде последовательности грамматических синтагм слов
- Строит поисковое представление предикатно-актантную структуры предложения и его «скелета» на основе грамматических синтагм слов
- Строит поисковое представление структуры словосочетаний с их контекстным окружением на основе грамматических синтагм слов

Концептуальный анализ текстов

- Выделение наименований понятий (сущностей) выполняется на этапе концептуального анализа текстов.
- Концептуальный анализ текстов - это лингвистическая процедура, обеспечивающая выявления их понятийного (концептуального) состава, формализации наименований понятий и установления смысловых связей между ними.
- Исходными данными для концептуального анализа являются результаты семантико-синтаксического анализа текстов
- Результатами работы концептуального анализа является система понятий анализируемого текста и их смысловые взаимосвязи

Методы концептуального анализа текстов

- Концептуальный анализ с контролем по эталонному концептуальному словарю (ЭКС) объемом 1.5 млн. наименований понятий.
- Концептуальный анализ текстов на основе “логической шкалы” словаря ЭКС.
- Концептуальный анализ текстов на основе синтаксических структур словаря ЭКС.
- Концептуальный анализ на основе обобщенных грамматических синтагм.
- Гибридный метод выявления наименований понятий из текстов СМИ.

Концептуальный анализ текстов с контролем по словарю ЭКС

1.Идея алгоритма: если некоторому отрезку текста соответствует в эталонном словаре хотя бы одно наименование понятия, имеющее такую же длину и такую же синтаксическую структуру, то этот отрезок текста с большой вероятностью также является наименованием понятия.

2.Условие: текст должен быть разделен на синтаксические предложения и каждое предложение должно быть разделено на всевозможные фрагменты последовательностей контактно расположенных слов

Концептуальный анализ текстов с контролем по словарю ЭКС

Шаг 1. Членение входного текста на предложения;

Шаг 2. Морфологический анализ текста;

Шаг 3. Пословная нормализация текста;

Шаг 4. Членение предложений текста на отдельные слова и отрезки текста длиной от 1-х до 5-ти слов;

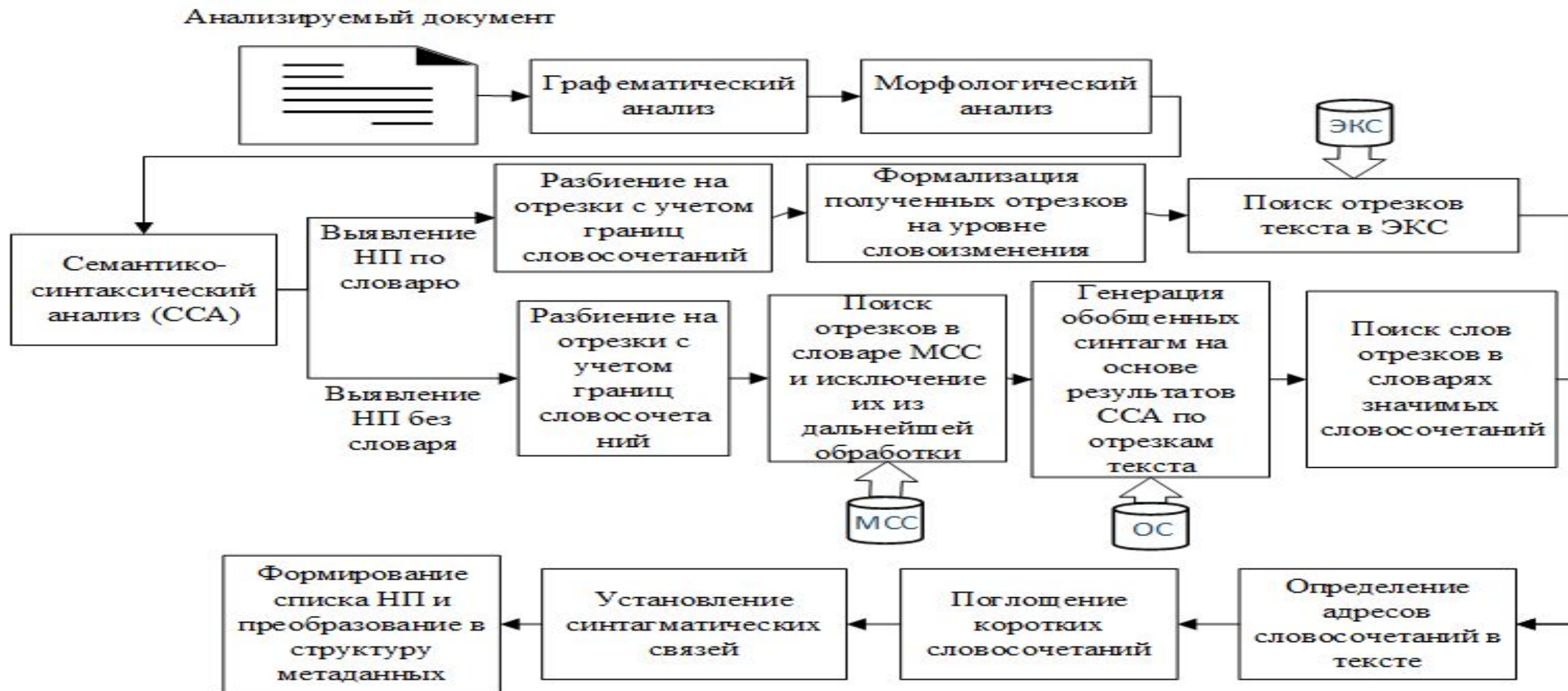
Шаг 5. Формирование поисковых образов слов и словосочетаний;

Шаг 6. Поиск в словаре ЭКС нормализованных текстовых фрагментов.

Шаг 7. Исключение из результатов поиска слов и словосочетаний, которые на одних и тех же отрезках текста входят в состав других, более длинных словосочетаний.

Шаг 8. Преобразование полученных результатов в структуру метаданных.

Гибридный метод выявления наименований понятий



Установления смысловой близости документов

- Преобразование текстового представления в его **формализованное смысловое представление** дает возможность сопоставления текстов по их смысловому содержанию.
- Такое сопоставление смыслового содержания текстов, обеспечивающее **выявление идентичных по смыслу фрагментов текстов** должно удовлетворять следующим условиям:
 1. В двух текстах должна быть **пересекающаяся совокупность наименований понятий**. Число понятий этой совокупности должно быть равно или превышать число наименований понятий, входящих в состав единичного высказывания.
 2. В двух таких текстах должны быть фрагменты, в которых **концентрация пересекающихся наименований понятий** превышает пороговое значение. Эти фрагменты должны иметь соизмеримые размеры.
 3. Эти фрагменты текстов должны быть **сходными по составу**

Технология составления словарей по корпусу текстов документов

1. Формирование репрезентативного тематического корпуса текстов
2. Морфологический анализ слов корпуса текстов
3. Семантико-синтаксического анализ корпуса текстов
4. Концептуальный анализ корпуса текстов и автоматическое выделение наименований понятий
5. Автоматическое приведение наименований понятий к их канонической форме
6. Формирование частотного словаря наименований понятий
7. Лингвистический анализ частотного словаря наименований понятий
8. Формирование машинного словаря и включение его в БД декларативных средств

Технология первичной обработки текстов документов

- Формально-логический контроль текста
- Морфологический анализ
- Синтаксический анализ
- Концептуальный анализ
- Автоматическое смысловое структурирование документов на предложения и сверхфразовые единства (последовательность контекстно-связанных предложений)
- Формирование различных форм формализованного представления текста
- Автоматическая классификация текста
- Формирование смысловой структуры текста в виде графа понятий и их отношений

Задачи автоматической обработки текстов при создании терминологических онтологий

- Выявление системы понятий в предметных областях на основе методов семантико-синтаксического и концептуального анализа текстов
- Установление системы парадигматических и синтагматических отношений между понятиями в предметной области
 - Отношений синонимии
 - Родовидовых отношений
 - Отношений типа часть – целое
 - Ассоциативных отношений

Задачи автоматической обработки текстов при создании фактологических онтологий

- Выявление системы фактов в предметных областях на основе методов семантико-синтаксического и концептуального анализа текстов
- Установление системы отношений между выявленными фактами и разработка методов установления их достоверности и непротиворечивости в предметной области

Инструменты для автоматизированного создания онтологий

- Процедуры семантико-синтаксического и концептуального анализа текстов
- Процедуры нормализации и отождествления различных форм слов
- Декларативные средства для смысловой обработки текстов
 - Эталонный политематический концептуальный словарь - 1.5 млн.
 - Словарь смысловых связей слов (словарь 4с) – 88 тыс.
 - Фразеологический словарь синонимичных конструкций – 400 тыс.
 - Грамматические таблицы для семантико-синтаксического анализа и концептуального анализа текстов – общий объем 1.2 млн. элементов

Машинный фразеологический перевод

1. Системы машинного перевода текстов с одних естественных языков на другие **моделируют работу человека-переводчика.**
2. Их эффективность зависит прежде всего от того, в какой степени в них учитываются **объективные законы функционирования языка и мышления.** А эти законы пока еще недостаточно изучены.
3. Поэтому, решая задачу машинного перевода, необходимо учитывать **богатый опыт межнационального общения и опыт переводческой деятельности, накопленный человечеством.**
4. Этот опыт свидетельствует о том, что в процессе перевода текстов в качестве основных единиц смысла выступают прежде всего не отдельные слова, а **фразеологические словосочетания, выражающие понятия.**
5. Именно понятия являются теми **элементарными мыслительными образами,** используя которые можно строить более **сложные мыслительные образы,** соответствующие переводимому тексту.

Концепция фразеологического машинного перевода (ФМП)

При создании систем ФМП необходимо руководствоваться следующими принципами:

1. Основными единицами языка и речи, которые прежде всего следует включать в машинный словарь, должны быть **фразеологические единицы (словосочетания, фразы)**.

Отдельные слова также могут включаться в словарь, но они должны использоваться только в тех случаях, когда не удастся осуществить перевод, опираясь только на фразеологические единицы.

2. Наряду с фразеологическими единицами, состоящими из непрерывных последовательностей слов, в системах машинного перевода могут использоваться и так называемые **"речевые модели" - фразеологические единицы с "пустыми местами"**, которые могут заполняться различными словами и словосочетаниями, порождая осмысленные отрезки речи.

Концепция ФМП

3. Реальные тексты, независимо от их принадлежности к той или иной тематической области, обычно бывают **политематическими**, если они имеют достаточно большой объем. И отличаются они друг от друга **не столько словарным составом, сколько распределениями вероятностей появления в них различных слов и словосочетаний** из общенационального словарного фонда. Поэтому машинный словарь, предназначенный для перевода текстов даже только из одной тематической области, должен быть политематическим, а для перевода текстов из различных предметных областей - тем более.

4. Для систем фразеологического перевода **необходимы машинные словари большого объема**. Такие словари должны создаваться прежде всего **на основе автоматизированной обработки двуязычных текстов**, являющихся переводами друг друга, и в процессе функционирования систем перевода.

Концепция ФМП

5. Наряду с основным политематическим словарем большого объема, в системах фразеологического машинного перевода целесообразно использовать также **набор небольших по объему дополнительных тематических словарей**. Дополнительные словари должны содержать только ту информацию, которая отсутствует в основном словаре.

6. Основным **средством разрешения полисемии** (многозначности) слов в системах фразеологического перевода является их использование в составе фразеологических словосочетаний. Дополнительным - **аппарат дополнительных тематических словарей**, где для каждого многозначного слова или словосочетания указывается его приоритетный переводной эквивалент, специфичный для рассматриваемой предметной области.

Концепция ФМП

7. Большую роль в системах фразеологического машинного перевода текстов могут играть **процедуры морфологического и синтаксического анализа и синтеза текстов, построенные на основе принципа аналогии**. Эти процедуры позволяют отказаться от хранения в словарях большого объема грамматической информации и порождать ее по мере необходимости автоматически, в процессе перевода. Они делают систему перевода открытой - способной обрабатывать тексты с "новой" лексикой.
8. Наряду с переводом текстов в автоматическом режиме, в системах фразеологического машинного перевода целесообразно предусмотреть и **интерактивный режим их работы**. В этом режиме пользователь должен иметь возможность вмешиваться в процесс перевода и **настраивать дополнительные машинные словари на тематику переводимых текстов**.



ИСТОРИЯ РАЗРАБОТКИ системы ФМП RETRANS

- 1957-1975 гг. - Научные исследования и разработка концепции ФМП**
- 1975-1993 гг.- Разработка базовых процедур и создание словарей**
- 1994 - Первая версия системы RETRANS (Платформа DOS)**
- 1998 – Сервер-переводчик для ВИНТИ (Платформа UNIX)**
- 2001 - Вторая версия системы RETRANS (Платформа Win-32)**
- 2002 – RETRANS VISTA (первая коммерческая версия)**
- 2004 – Многопользовательская система для ФНС России**
- 2005 – SU PETRANS 2005 ML-2 (персональная коммерческая версия)**
- 2006 - RETRANS 2005 Server (сервер-переводчик Интернет-сайтов)**
- 2007 - MU PETRANS 2005 ML-2 (многопользовательская коммерческая версия)**
- 2008 - SU PETRANS 2008 ML-2 (персональная коммерческая версия)**
- 2008 - MU PETRANS 2008 ML-2 (многопользовательская коммерческая версия)**
- 2008 -Portal MU RETRANS Server (портальный многопользовательский сервер-переводчик).**



ИСТОРИЯ РАЗРАБОТКИ системы ФМП RETRANS

- 1957-1975 гг. - Научные исследования и разработка концепции ФМП**
- 1975-1993 гг.- Разработка базовых процедур и создание словарей**
- 1994 - Первая версия системы RETRANS (Платформа DOS)**
- 1998 – Сервер-переводчик для ВИНТИ (Платформа UNIX)**
- 2001 - Вторая версия системы RETRANS (Платформа Win-32)**
- 2002 – RETRANS VISTA (первая коммерческая версия)**
- 2004 – Многопользовательская система для ФНС России**
- 2005 – SU PETRANS 2005 ML-2 (персональная коммерческая версия)**
- 2006 - RETRANS 2005 Server (сервер-переводчик Интернет-сайтов)**
- 2007 - MU PETRANS 2005 ML-2 (многопользовательская коммерческая версия)**
- 2008 - SU PETRANS 2008 ML-2 (персональная коммерческая версия)**
- 2008 - MU PETRANS 2008 ML-2 (многопользовательская коммерческая версия)**
- 2008 -Portal MU RETRANS Server (портальный многопользовательский сервер-переводчик).**



РЕАЛИЗОВАННЫЕ ПРОЕКТЫ



Администрация Президента РФ

Многопользовательская система
фразеологического машинного перевода
RETRANS GOV

Разработка многопользовательского программного
комплекса для выпол-нения автоматизированного
перевода англоязычных документов для сотрудников
ЦА Администрации Президента РФ на основе
использова-
ния системы MU RERTRANS ML-2



РЕАЛИЗОВАННЫЕ ПРОЕКТЫ



**Совет Федерации
ФС РФ**



**Государственная
Дума РФ**

**Подсистема машинного перевода Единого
Интернет портала Феде-рального Собрания
Российской Федерации**

Разработка и поддержка системы публикаций
англоязычных инфор-мационных материалов для
разме-щения на web-сайте “Парламен-тский
портал”

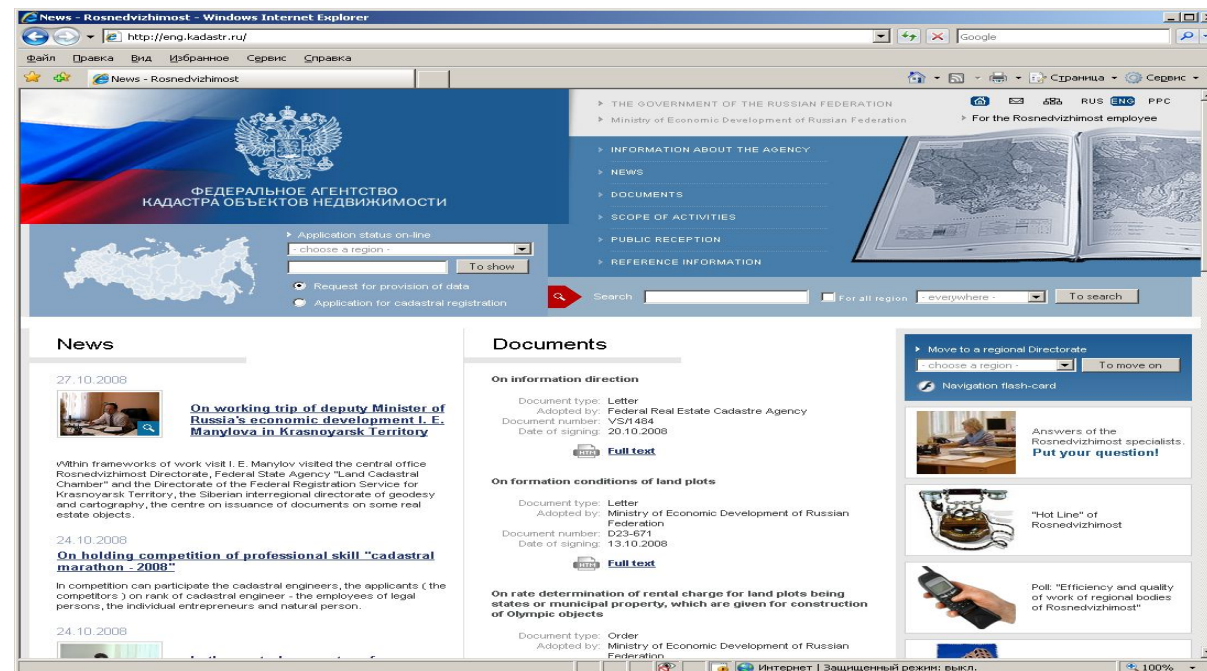


РЕАЛИЗОВАННЫЕ ПРОЕКТЫ

Разработка сервера-переводчика для англоязычной версии Интернет-сайта **РОСНЕДВИЖИМОСТЬ**



Федеральное агентство
РОСНЕДВИЖИМОСТЬ



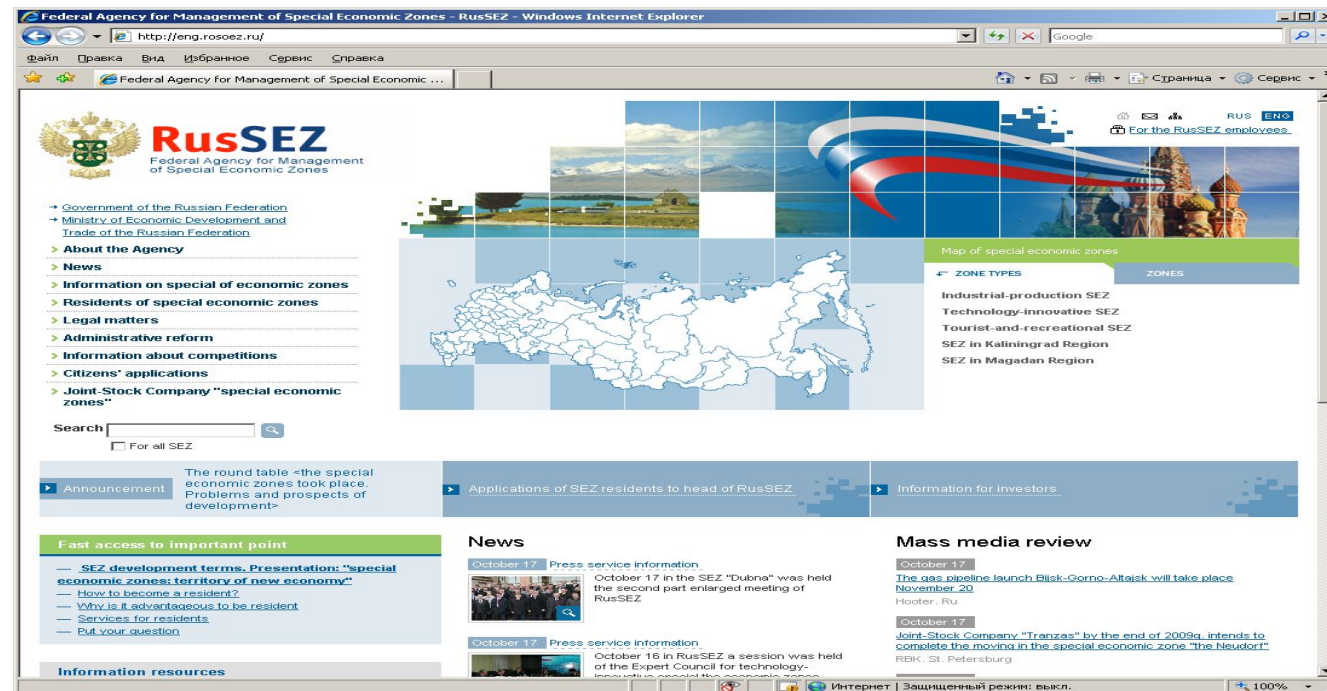


РЕАЛИЗОВАННЫЕ ПРОЕКТЫ

Разработка сервера-переводчика для англоязычной версии Интернет-сайта РосОЭЗ



Федеральное
агентство РосОЭЗ



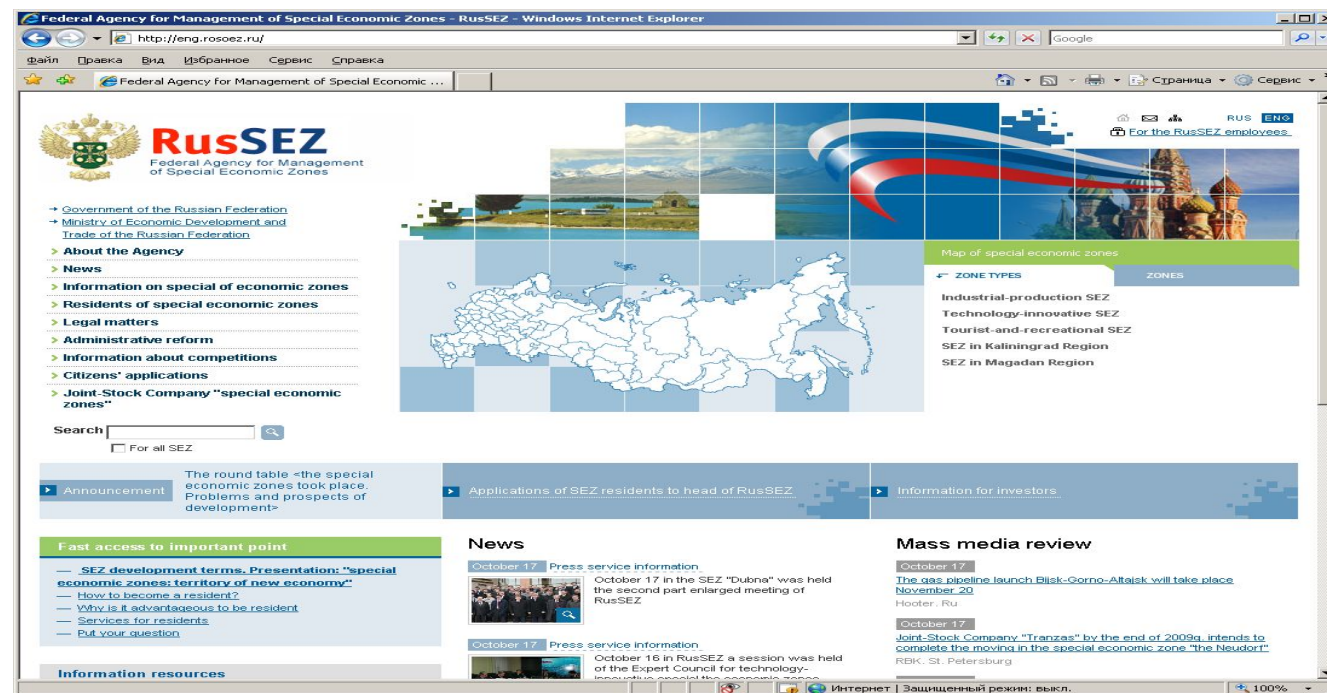


РЕАЛИЗОВАННЫЕ ПРОЕКТЫ

Разработка сервера-переводчика для англоязычной версии Интернет-сайта РосОЭЗ



Федеральное
агентство РосОЭЗ





**Смысловая
поисково-аналитическая
система авиационной
отрасли
«СПАС-Авиа»**

ФГУП «ГосНИИАС», 2012-2018 г.



**Комплекс
лингвистического
программного обеспечения
MetaFraz R10**

MetaFraz, 2012-2018 г.



Программно- лингвистическая платформа MetaFraz R10

MetaFraz, 2016-2018 г.



**Система
семантической
обработки текстов
MetaFraz R10
(MF TEXT ANALYST R10)**

MetaFraz, 2016-2018 г.

СПАСИБО ЗА ВНИМАНИЕ

