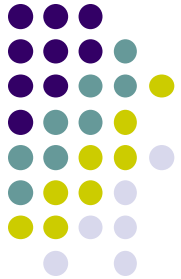


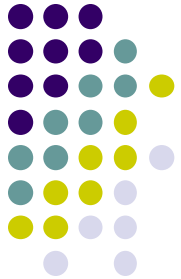
АВТОМАТИЧЕСКАЯ ОБРАБОТКА ТЕКСТОВ: ЗАДАЧИ, ПОДХОДЫ, РЕСУРСЫ

Большакова Елена Игоревна

МГУ имени М.В.Ломоносова, ф-т ВМК

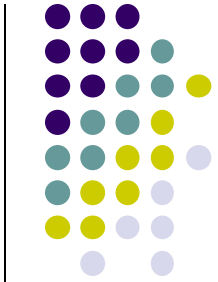
Летняя школа по АОТиАД, 24 июля 2017

СОДЕРЖАНИЕ



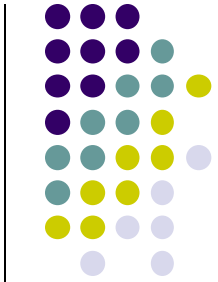
1. Компьютерная лингвистика (КЛ) и автоматическая обработка текстов (АОТ): истоки, междисциплинарность, задачи
2. Особенности естественного языка (ЕЯ)
   - уровни и единицы языка и текста
   - неоднозначность языковых знаков
3. Моделирование в КЛ
4. Этапы обработки текста на ЕЯ
5. Лингвистические ресурсы
6. Подходы к построению систем обработки ЕЯ
7. Прикладные задачи АОТ и КЛ

КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА: ИСТОКИ



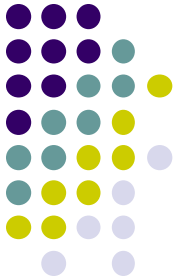
- Начало работ – 50-е годы,
  Потребности практики: машинный перевод
- Название научной области:
  - Автоматическая обработка тестов на *естественном языке* (ЕЯ) – *Natural Language Processing*
  - Вычислительная/ Компьютерная лингвистика
    Computational Linguistics
- Междисциплинарная научная область:
  - Лингвистика
  - Математика
  - Информатика (*Computer Science*)
  - Искусственный интеллект (*Artificial Intelligence*)

КЛ: ЛИНГВИСТИКА И МАТЕМАТИКА



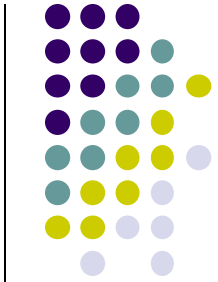
- Общая лингвистика
  - Фонология (звуки речи)
  - Морфология (структура и форма слов ЕЯ)
  - Синтаксис (структура и функции предложений)
  - Семантика (смысл языковых высказываний)
  - Прагматика (значение высказываний)
- Математическая лингвистика (область математики)
  - *Теория формальных языков и грамматик* возникла из *порождающих грамматик* Н.Хомского (50-е гг.) для анализа синтаксических структур ЕЯ
  - *Квантитативная (статистическая)* лингвистика: изучение языка/речи количественными методами

КЛ: ИНФОРМАТИКА и ИИСКУССТВЕННЫЙ ИНТЕЛЕКТ



- Информатика (*Computer Science*)
  - Общая методология с КЛ – построение компьютерных программ
  - Методы трансляции языков программирования (ЯП) – алгоритмы из теории формальных языков
- Искусственный интеллект:
  - Задача: компьютерное моделирование интеллектуальных функций
  - Пересечение с КЛ: обработка ЕЯ – интеллектуальная функция
  - Методы моделирования: *эвристические*

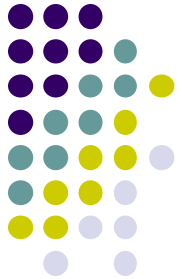
ОСНОВНАЯ ЗАДАЧА КЛ



- Цель – построение систем для автоматической обработки информации, представленной на *ЕЯ*
- Задача: Разработка *лингвистических процессоров* для различных прикладных систем
  - Лингвистический процессор: модуль или вся система
  - Основа процессора – формальная модель текста/языка

  Проблемы связаны со сложностью языка
- Естественный язык – сложная система знаков (звуковых и письменных), возникшая в процессе человеческой деятельности как средство общения
- Функции ЕЯ: коммуникация, мышление, познание и сохранение знаний

ОСОБЕННОСТИ ЕЯ КАК ЗНАКОВОЙ СИСТЕМЫ

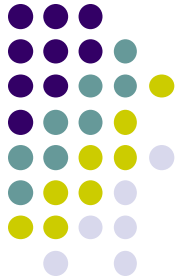


Сложная комбинаторная система знаков:

- Постоянная изменчивость
- Несколько сот тысяч языковых знаков
- Многоуровневость: каждый уровень (подсистема) – правила сочетания *единиц* (знаков) этого уровня
- Взаимосвязь, иерархия уровней
- Избыточность ЕЯ (но и универсальность)
- Многозначность, неопределенность смысла знаков

Невозможность один раз и навсегда создать лингвистический процессор

МНОГОУРОВНЕВОСТЬ ЕЯ

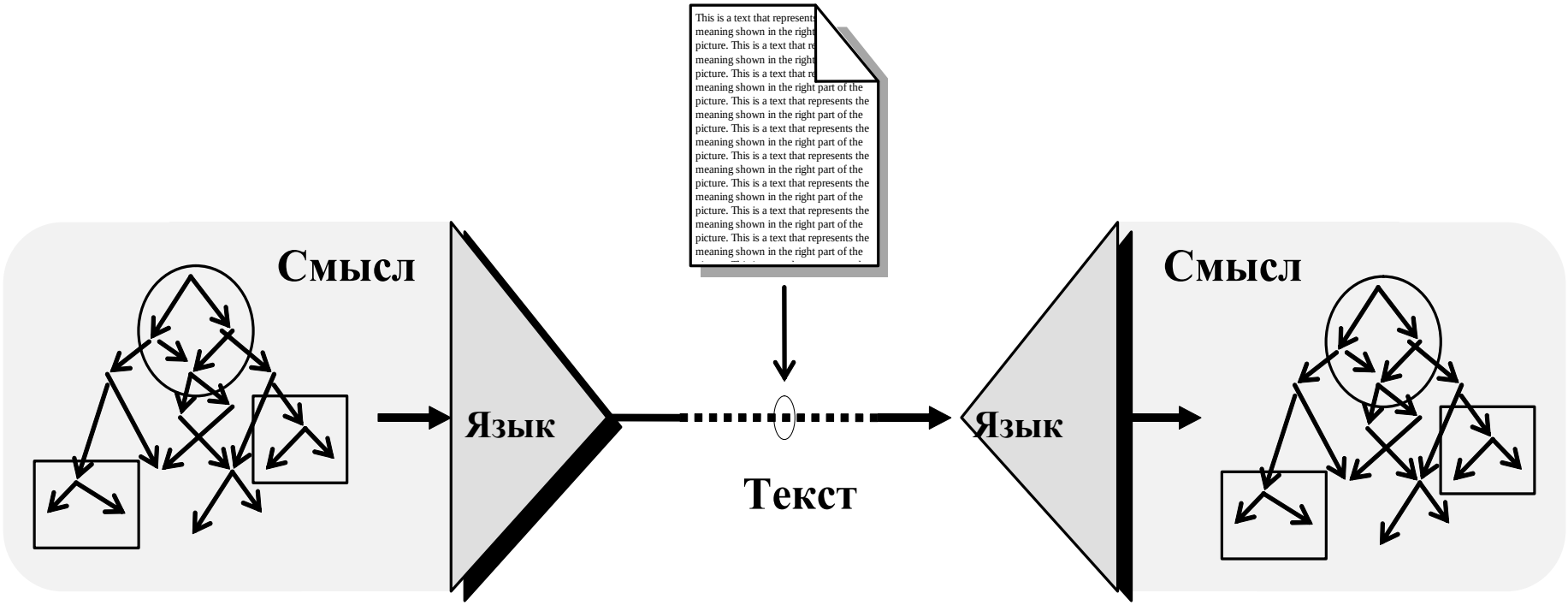


Уровни ЕЯ взаимосвязаны; иерархия уровней:
разложимость единиц одного уровня на меньшие)

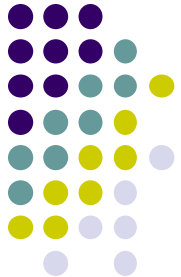
- **фонологический**: звуки (*фонемы*) / буквы
- **морфологический**: слова (*словоформы*): *листом*
  - подуровень *морфем* (корень, суффикс…)*: по-стро-ен*
- **лексический**: *лексемы* (*лексикон*)
  лексема – совокупность *словоформ* слова
  например: *лист*, *листа*, *листу*, *листе, ..*
- **синтаксический**: предложения (фразы)
  - подуровень *словосочетаний* : *синий цвет, смотрю кино, чай с сахаром*
  - надуровень *сверхфразовых единств* (≈ абзацев)
- **семантический** (смысловой) : *семы*
- **дискурсивный** : структуры связного текста

ЯЗЫК КАК ПРЕОБРАЗОВАТЕЛЬ СМЫСЛ ⇔ ТЕКСТ

- Центральный объект – *текст*, линейность текста
- Текст составлен из *единиц* разного *уровня*
- Единицы: незначащие и значащие (языковые знаки)



ЯЗЫК и РЕЧЬ (ТЕКСТ)



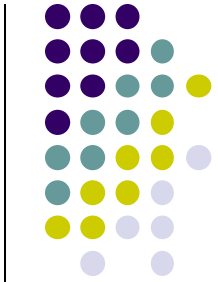
Разграничение в лингвистике:

- Язык: система знаков ЕЯ
- Речь (устная, письменная): линейная последовательность знаков, построенная в процессе общения, в соответствии с принятыми правилами

Единицы:

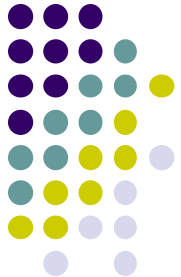
- Языка:
  - фонемы / графемы(буквы)
  - морфемы
  - лексемы (слова)
- Речи / текста:
  - буквы, морфы,словоформы (словоформы)
  - словосочетания
  - предложения (фразы) …

ДРУГИЕ СЛОЖНОСТИ ЕЯ



- Нестандартная сочетаемость (*синтактика*) единиц ЕЯ на всех уровнях, например, лексическая:
  крепкий чай, но не *сильный чай* (*strong tea*)
- Неоднозначность языковых единиц
  - *Полисемия* – многозначность языковой единицы, например, для слова *земля*:
    Земля, суша, почва, страна, территория
  - *Синонимия* – совпадение единиц по основному смыслу: синонимия слов: *горячий – жаркий*
    синонимия предлогов: *о поездке – про поездку*
    синонимия приставок, суффиксов, союзов и др.
  - *Омонимия* – совпадение по форме двух или более языковых единиц (отличие: нет смысловой связи между совпавшими по форме единицами)

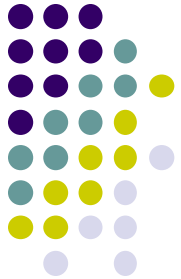
ЕЯ : ОМОНИМИЯ



Звуковое совпадение или совпадение на письме двух разных по смыслу единиц. Наиболее частые виды:

- *Лексическая омонимия* – одинаково звучащие/пишущиеся слова, не имеющие общих элементов смысла,
  например: *рожа* – лицо и вид болезни.
- *Морфологическая омонимия* – совпадение форм одного и того же слова (лексемы) : *лист* (имен. и винит. Падеж)
- *Лексико-морфологическая омонимия* – совпадение словоформ двух разных лексем, например:
  стих – глагол в единств. числе мужского рода и существительное в единств. числе, именит. падеже
- *Синтаксическая омонимия* – неоднозначность синтаксической структуры (и соответствующего смысла):
  Студенты из Львова поехали в Киев
  Flying planes can be dangerous (пример Хомского)

МОДЕЛИРОВАНИЕ в КЛ

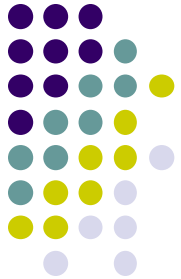


Лингвистический процесор опирается на модель языка, которая должна обладать структурным и\или функциональным подобием

Особенности моделей КЛ (отличие от лингвистических):

- Формальность и алгоритмизируемость
- Функциональность: воспроизведение функций языка, а не моделирование языковой деятельности человека
- Общность модели, т.е. покрытие ею довольно большого множества текстов
- Экспериментальная обоснованность (*Evaluation*)
- Ориентация на конкретные прикладные задачи КЛ
- Опора на те или иные лингвистические ресурсы как обязательную составляющую модели

ПРИКЛАДНЫЕ ЗАДАЧИ КЛ



Модель ЕЯ выбирается для конкретного приложения:

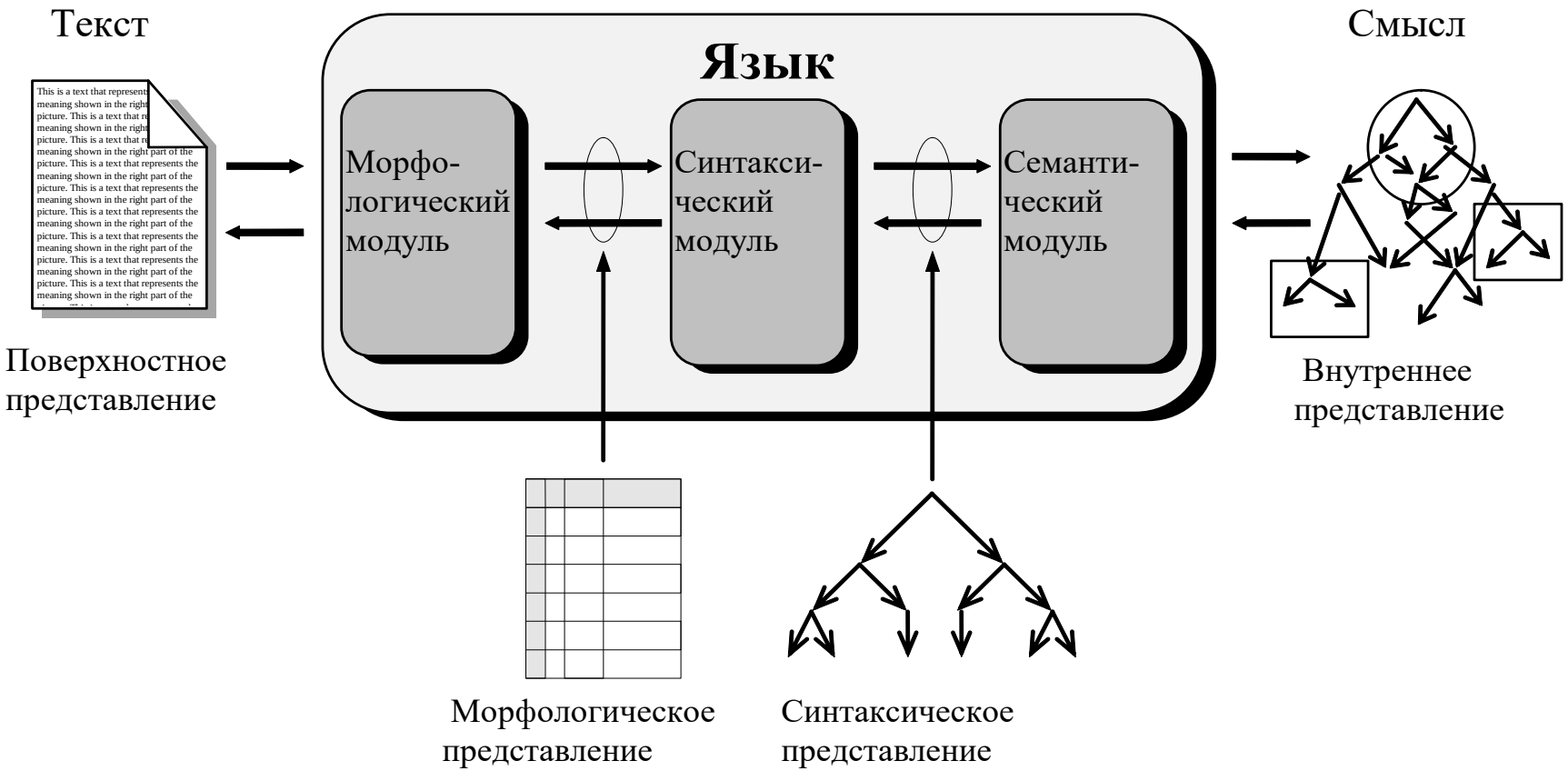
- Машинный перевод
- Информационный поиск
- Реферирование и аннотирование текстов
- Автоматизация создания и редактирования текстов
- Генерация текстов на ЕЯ
- Формирование ответов на вопросы
- Организация диалога (общения) на ЕЯ
- Распознавание и синтез звучащей речи

Text Mining:

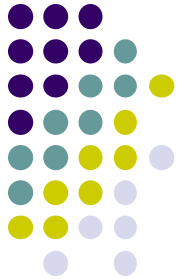
- Извлечение информации из текстов
- Классификация и кластеризация текстов
- Извлечение терминов и ключевых слов
- Анализ мнений и оценка тональности текстов

ЭТАПЫ ОБРАБОТКИ ТЕКСТА

В общем случае: лингвистический процессор – многоэтапный преобразователь (два направления: анализ и синтез)



УРОВНИ АНАЛИЗА ТЕКСТА

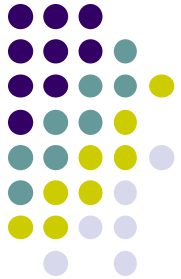


Уровни (этапы) анализа ~ Уровни языка

- Графематический анализ: *сегментация*
- Морфологический анализ
  - Постморфологический анализ: *разрешение морфологической омонимии*
    (часто – функция морфологического процессора)
  - Предсинтаксис: сегментация текста на предложения , выделение словосочетаний
- Синтаксический анализ предложений
- Семантический и дискурсивный анализ

❖ *глубина обработки* текста: количество уровней

МОРФОЛОГИЧЕСКИЙ АНАЛИЗ



Вход: словоформа текста ЕЯ

Виды морфологической обработки:

- Лемматизация (синоним: *нормализация)*

  Выход: *лемма* = словарная/стандартная форма слова

  красивее → красивый, лег → лечь

- Стемминг

  Выход: основа/псевдооснова слова

  водных → водн / вод

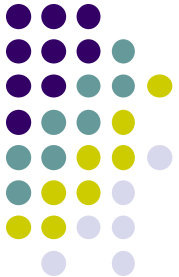
- Полный морфоанализ

  Выход: лемма **+** морфол. характеристики (*теги*)

  водных → водный + прилагательное,
  множ.число, родит.падеж

! Возможно несколько вариантов анализа (омонимия)

РАЗРЕШЕНИЕ МОРФОЛОГИЧЕСКОЙ ОМОНИМИИ

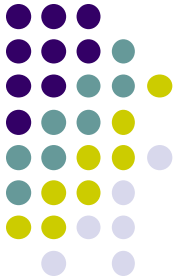


Разрешение /Снятие морфологической омонимии (*Morphological Disambiguation*) – устранение морфологической многозначности: *стали, зала*

- выбор правильной леммы
- уточнение морфологических характеристик

- Снятие – предсинтаксический этап: может быть встроен в морфопроцессор или реализован отдельно
- Основные методы:
  - Лингвистические правила, например: удаление всех омонимов слова с падежами, не соответствующими возможным падежам предшествующего предлога: *у зала* (возможен предл., но не именит. падеж)
  - Машинное обучение

МОРФОАНАЛИЗАТОРЫ ДЛЯ РУССКОГО ЯЗЫКА



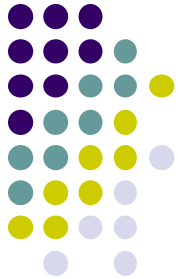
Свободный доступ, полный морфоанализ

- *Mystem* компании Яндекс (исполняемый модуль)
  http://company.yandex.ru/technology/mystem
  - есть сегментация и контекстное снятие омонимии
- Морфопроцессор АОТ проекта Диалинг : *www.aot.ru*
  - есть онлайн-интерфейс, открытый код на C++
- Морфопроцессор *Pymorphy2*

https://pymorphy2.readthedocs.org/en/0.2/user/index.html

  - слабая сегментация, удобен для *Python*
- Модуль *TreeTagger* –сегментация, снятие омонимии
  http://corpus.leeds.ac.uk/mocky/

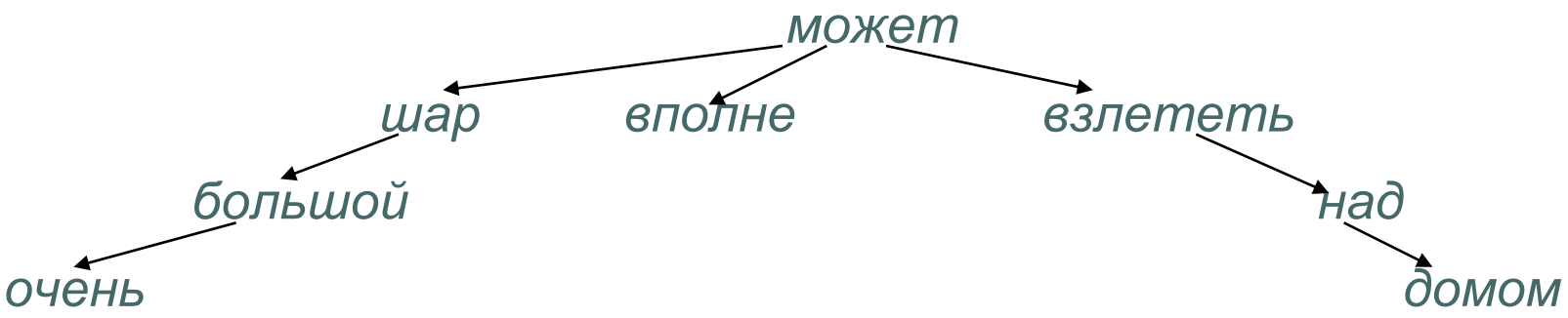
СИНТАКСИЧЕСКИЙ АНАЛИЗ



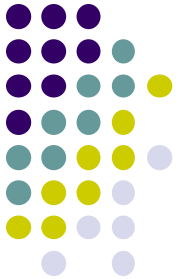
- На входе: предложение ЕЯ +
  результат морфологического анализа
- На выходе: ***синтаксическое дерево***
  (структура) предложения
- Модели синтаксической структуры предложения:
  - *деревья зависимостей/ подчинения*
  - *деревья составляющих*
- Модели СА отличаются:
  – *синтаксическими единицами* и
  – *синтаксическими связями* между ними
- Возникли соответственно в Европе и Америке – для ЕЯ с разным синтаксисом

МОДЕЛЬ СИНТАКСИСА: ДЕРЕВЬЯ ЗАВИСИМОСТЕЙ

- Основа – **подчинительная связь** слов
- Дерево зависимостей (*dependency*) предложения:
  - ✓ узлы – слова (корень дерева – глагол, сказуемое)
  - ✓ дуги – подчинительная связь (зависимость)
- Особенность: дерево предложения должно быть дополнено информацией о линейной структуре (т.е. задан порядок слов)
- Пример дерева синтаксических зависимостей:



МОДЕЛЬ СИНТАКСИСА: ДЕРЕВЬЯ СОСТАВЛЯЮЩИХ



- Синтаксические единицы – *составляющие (constituents)*, т.е. отрезки текста , в том числе слова, словосочетания предложение в целом
- Могут вкладываться друг в друга, но не пересекаться
- Связь этих синтаксических единиц – отношение вложения – графически изображается как *дерево составляющих.*
- Грамматически правильная синтаксическая структура обычно фиксируется КС-грамматикой (по Хомскому).
- *КС-грамматика* для примера дерева составляющих:

  $S \rightarrow NP\ VP\ NP$ $NP \rightarrow N \mid A\ N \mid Det\ N \mid N\ PP$

  $VP \rightarrow V \mid AnV \mid Adv\ V \mid Adv\ AnV$ $AnV \rightarrow Aux\ V$

  $PP \rightarrow Prep\ NP$
- Нетерминалы – фактически метки-типы составляющих.

РАЗМЕЧЕННОЕ ДЕРЕВО СОСТАВЛЯЮЩИХ

Метки типы фраз:

S – предложение

NP – именная группа

VP – глагольная группа

AnV – аналитическая форма глагола

PP – предложная группа

Det – местоименное прилагательное

N – имя существительное

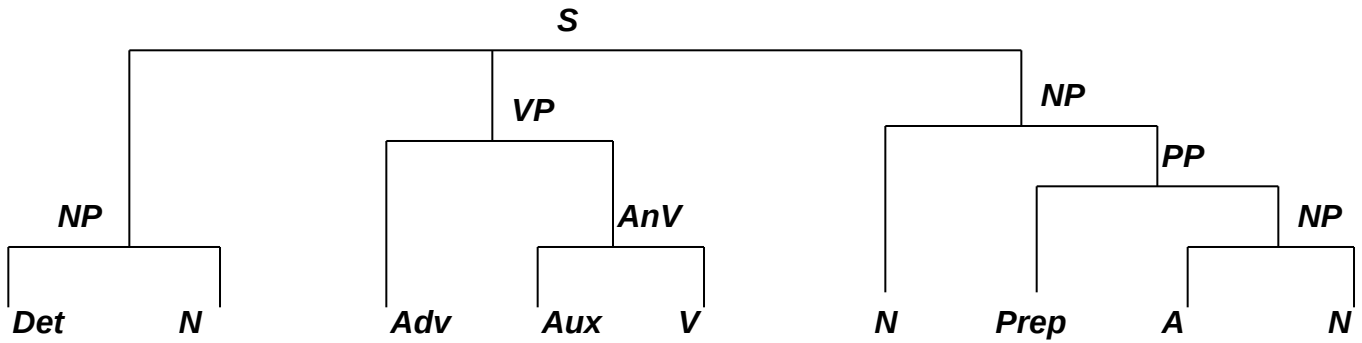
Adv – наречие

Aux – вспомогательный глагол

V – глагол

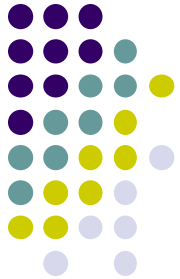
Prep – предлог

A – имя прилагательное



(Эти школьники) (скоро (будут писать)) (диктант (по (русскому языку)))

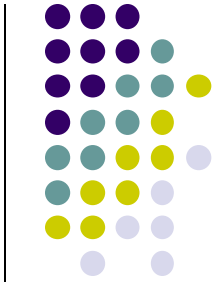
СИНТАКСИЧЕСКИЕ ПАРСЕРЫ ДЛЯ РУССКОГО ЯЗЫКА



Свободный доступ

- Модуль синтаксич. анализа *SynAn* проекта Диалинг :
  - *www.aot.ru* , открытый код на C++
  - гибридная модель (синтаксические группы)
  - онлайн-интерфейс: *http://www.aot.ru/demo/synt.html*
- *MaltParser*
  - на основе машинного обучения
  - предобработка текста:
    – морфологический анализ: *TreeTagger*
    – лемматизаторнеизвестных слов: *CSTLemma*
- ?

СЕМАНТИЧЕСКИЙ АНАЛИЗ



- На входе: синтаксическое дерево/деревья
- На выходе: семантическая структура

Модели представления смысла /семантики (свойства объектов, их отношения, состояния, действия) – на основе моделей ПЗ в ИИ: формулы исчисления предикатов или семантические сети

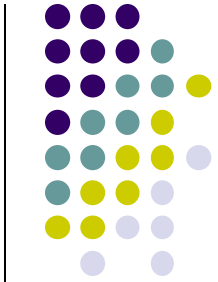
Локальный семантический анализ:

По синтаксическому дереву предложения построить его семантическую структуру: дерево/граф или формулу логики

Подзадачи:

- Определение семантики слов (и словосочетаний), включая разрешение их многозначности слов: *лук, оператор, дом*
- Установление *семантических* отношений между словами и словосочетаниями

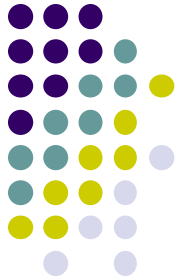
СЕМАНТИЧЕСКИЙ АНАЛИЗ ПРЕДЛОЖЕНИЯ



Простое предложение ЕЯ – *пропозиция*
(отдельная мысль, высказывание):
Федор переехал из Москвы в Питер.

- Семантический анализ опирается на предикатную, актантно-аргументную структуру предложения, выявляемую в ходе синтаксического анализа.
- Корень синтаксического дерева зависимостей – слово-предикат (сказуемое)
- Слова-предикаты имеют места для заполнения – *валентности*: *Подарить*: кто? (1) что? (2) кому? (3)
  Переехать: кто? (1) откуда? (2) куда? (3)
- Заполнители – актанты (по сути – аргументы предиката), это слова и словосочетания, т.о. в итоге формула логики
  Переехать (Федор, Москва, Питер)

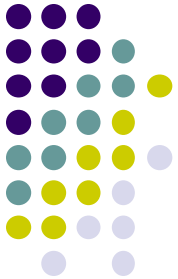
ДИСКУРСИВНЫЙ АНАЛИЗ



Дискурс (*фр.*: *discours* – речь); рассмотрение связного текста в целом, с учетом его назначения

- Пропозиции-предложения д. б. связаны между собой
- Анализируются:
  - Локальная связность предложений (*cohesion*):
    анафорические отсылки, лексические повторы –
    Я читала эту книгу. Книга (она) была интересной.
  - Глобальная связность (целостность, *coherence*):
    - *тематическая*
    - *композиционная/сюжетная* структура текста:
      рассказы, пьесы, сказки
      деловые письма, заявления, юридич. документы
      научные статьи, технические и патентные описания

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ



Лингвистические процессоры опираются на лингвистические ресурсы.

Источники информации о языке

Словарные (лексические) ресурсы:

- Компьютерные словари
- Тезаурусы, Онтологии

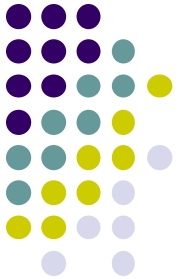
Текстовые ресурсы:

- Коллекции текстов (предметной области)
- Корпуса текстов

Ресурсы смешанного типа:

FrameNet – слова, предложения с лингв. данными

ЛЕКСИЧЕСКИЕ РЕСУРСЫ: СЛОВАРИ

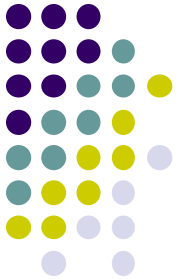


Компьютерные словари различаются:

- охватом лексики: общая/специальная
- единицами (словарными статьями):
  - словари синонимов: *бродить / шататься*
  - словари паронимов: *чужой / чуждый*
  - словари терминов предметной области
  - оценочных слов: *мерзкий, отличный*
  - словари (базы) устойчивых словосочетаний (*коллокаций*): *острая нехватка*

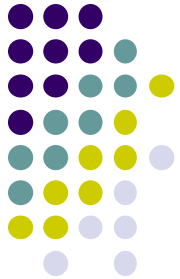
Интернет-ресурсы: *WiikiPedia,Wiki-словарь, DBPedia*

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ: ТЕЗАУРУСЫ И ОНТОЛОГИИ



- Тезаурус – семантический словарь
  - *РуТез* – информационно-поисковый тезаурус понятий из общественно-политической области; смысловые связи: синонимия, род-вид, ассоциация
- Онтология – формальное описание определенного набора понятий, сущностей
  - *WordNet* – лингвистич. онтология на базе англ. слов
    – Дж. Миллер (80е гг.), модель человеческой памяти
    – слова разбиты по частям речи, для каждой части речи выделены *синсеты* (синонимы) – понятия
    – версия 3.0 – 155 тыс. лексем, 117 тыс. синсетов
  - *EuroNet* – аналогичные лексические ресурсы для других европейских языков

ЛИНГВИСТИЧЕСКИЕ РЕСУРСЫ: КОРПУСА ТЕКСТОВ

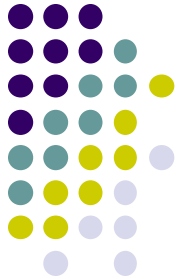


Применение: – построение словарей
– машинное обучение моделей ЕЯ и текста

- *Коллекция текстов:* набор объединенных по некоторому признаку текстов (напр.бнормативно-правовые документы)
- *Корпус текстов* – представительный массив текстов:
  - предназначен для решения конкретных лингвист. задач
  - обладает *лингвистической разметкой*: лексической, морфологической, синтаксической, дискурсивной
  - часто поддерживается спец. корпусным менеджером

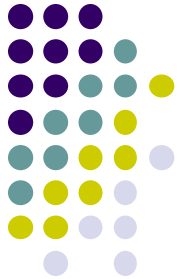
  РЯ: Национальный корпус русского языка (*НКРЯ*), *ГИКРЯ, OpenCorpora, RuTenTen, SynTagRus* (синтаксис)
- *Интернет-корпус:* корпус современной речи

ПОДХОДЫ К ПОСТРОЕНИЮ МОДУЛЕЙ КЛ



- Основанный на правилах, или инженерный:
  rule-based, knowledge-based
  - Модель – набор лингвистических правил
  - Правила создаются экспертами (лингвистами)
  - Обычно применяются специальные языки записи правил и соответствующие программные системы
- Основанный на машинном обучении
  - Виды обучения: – обучение с учителем (*supervised*)
    – обучение без учителя (*unsupervised*)
    – частичное обучения с учителем (*bootstrapping*)
  - Модель – машинный классификатор
  - Необходим размеченный текстовый корпус

МАШИННОЕ ОБУЧЕНИЕ



Вход: множество ситуаций (***C***) и реакций (***P***)
обучающая выборка: пары ***C-P***

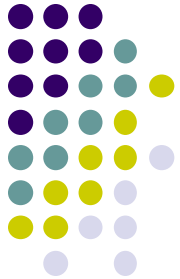
Задача: по обучающей выборке выявить зависимость между ***C*** и ***P***

Алгоритм обучения пытается найти эту зависимость

Выход: функция, ставящая в соответствие ***C*** определенную ***P***

\+ Не требуется работа эксперта
\+ Можно пробовать разные алгоритмы обучения
– Нужна большая, качественно размеченная обучающая выборка
– Сложно локализовать и исправить ошибку

ВЫБОР ПОДХОДА



- Нет экспертов ПО и словарных ресурсов
- Много размеченных данных и получение их дешево
- Необходимо быстрое построение приложения
- Не требуется лингв. интерпретация результатов

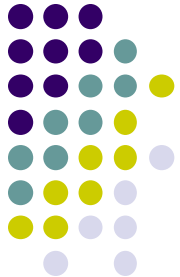
➔ **Машинное обучение**

- Есть эксперты и словарные ресурсы
- Мало размеченных данных
- Есть временные ресурсы
- Необходимо хорошее (и выше) качество работы

➔ **Инженерный подход**

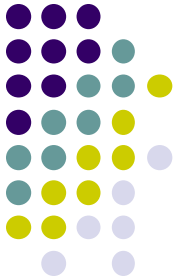
Качество работы: точность, полнота, F-мера

ВИДЫ МОДЕЛЕЙ В КЛ



- Многомодульные (*multi-component, pipelined*):
  модули относятся к разным уровням/этапам, и могут быть созданы в рамках разных подходов
- *Признаковая модель текста:* для обработки коллекций
  - признаки определены для каждого документа
  - *информационные признаки*: лингвистические, статистические, структурные характеристики текста
  - виды модели: *bag of words* (мешок слов), *векторная*
- *Статистическая языковая модель* (*Language Model*)
  - модель всего языка, строится по коллекции текстов
  - основана на статистике слов (или символов/букв) и их последовательностей – *N-грамм* (признаков)
  - отвечает на вопрос, насколько вероятно появление слова, если перед ним встречались конкретные слова

ПРИКЛАДНЫЕ ЗАДАЧИ КЛ



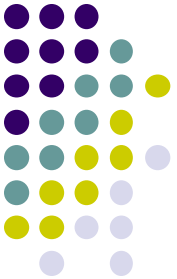
Традиционные направления:

- Машинный перевод
- Информационный поиск
- Реферирование и аннотирование текстов
- Автоматизация создания и редактирования текстов
- Генерация текстов на ЕЯ
- Формирование ответов на вопросы
- Организация диалога, чат-боты
- Распознавание и синтез звучащей речи

Text Mining:

- Извлечение информации из текстов
- Классификация и кластеризация текстов
- Извлечение терминов и ключевых слов
- Анализ мнений и оценка тональности текстов

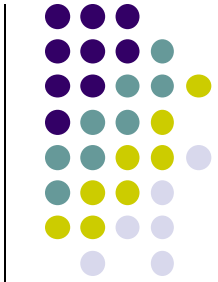
ПРИКЛАДНЫЕ ЗАДАЧИ КЛ: МАШИННЫЙ ПЕРЕВОД



Начало исследований - 50-е годы 20-го века

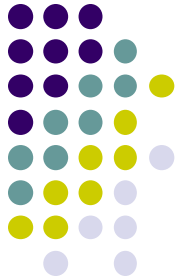
- Джоржтаунский эксперимент, 1954 г.: автоматический перевод с русского на английский, словарь – 250 слов
- Первые работы в России: 1955 г., словарь – 2300 слов; перевод с английского на русский
- Простейшая лингвист. модель: *пословный* перевод
- Неравномерность развития работ по МП (приостановка финансирования исследований в 60-е годы)
- 50-60 гг. – двуязычные системы, *пословный* и *пословно-пооборотный* перевод (приемлемое качество для родственных языков)

МАШИННЫЙ ПЕРЕВОД: ПОКОЛЕНИЯ СИСТЕМ



- 60-70 гг. – *пофразный* перевод,
  стратегия АНАЛИЗ ⇒ ТРАНСФЕР ⇒ СИНТЕЗ
  – пред- и пост-редактирование человеком
  – появление промышленных систем
- 70-80 гг., экстенсивное развитие: *многоязычные* системы
  - *ЭТАП* (СССР): лингв.модель ЕЯ «Смысл⇔Текст»,
    франц./англ.русский перевод научно-технич. текстов
- 80-90 гг. – многоязычные системы,
  – опора на лексические и терминологические БД
  – использование *интерлингвы* – языка-посредника
- 90-2000 гг. – применение статистики, корпусов текстов:
  статистическая трансляция
- ~ с 2010 гг. – машинное обучение на *нейронных сетях*

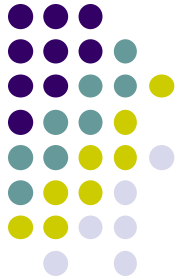
ПРИКЛАДНЫЕ ЗАДАЧИ КЛ: ИНФОРМАЦИОННЫЙ ПОИСК



- Поиск в коллекциях текст. документов – с 50-х гг.
  - *Поисковый образ* документа – *ключевые слова* (отражают основное содержание документа)
  - Поиск документа по запросу в виде набора ключ. слов
  - Результат поиска – *релевантные* документы
  - *Индексирование* документа, т.е. выделение ключевых слов и словосочетаний, выполнялось вручную

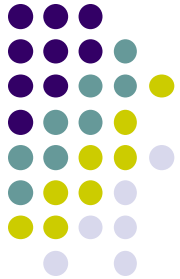
  Применяется в соврем. корпоративных инф. системах
- Полнотекстовый поиск – с 90-х гг. (в сети Интернет)
  - *Автоматическое индексирование* текстов
  - Применение *векторной модели* текста (*bag of words* : набор знаменательных слов текста с их частотами)

ИНФОРМАЦИОННЫЙ ПОИСК: СМЕЖНЫЕ ЗАДАЧИ



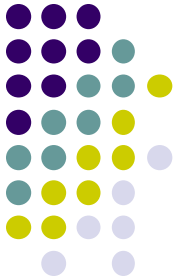
- Классификация текстов – отнесение к классам с заданными свойствами/параметрами
- Рубрицирование текстов – классификация, соотнесение с иерархической системой классов
- Кластеризация текстов – создание подмножеств тематически близких документов
- Построение *вторичных документов*:
  - Реферирование текста – построение краткого реферата для одного или нескольких текстов
  - Аннотирование текста – краткое описание содержания текста (упрощенно: список ключевых слов)

ПРИМЕНЕНИЕ КЛАССИФИКАЦИИ и КЛАСТЕРИЗАЦИИ



- Упорядочивание и навигация по набору документов
  - составление интернет-каталогов
- Информационный поиск
  - ограничение области поиска
  - «интеллектуальная» группировка результатов
- Фильтрация потока документов
  - фильтрация спама
  - выявление «искусственных» текстов (боты)
  - определение дубликатов документов
- Персонализированный подбор информации
  - контекстная реклама
  - новости об определенном событии и т.п.

ПРИКЛАДНЫЕ ЗАДАЧИ: *QUESTION ANSWERING*

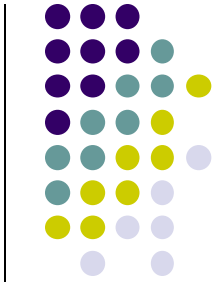


Ответы на вопросы –
сравнительно новая задача, актуальная

(но и забытое старое направление ИИ, 70 гг.)

- Нужен не документ или *сниппет*,
  а ответ на конкретный вопрос , например:
  Кто придумал вилку? ⇒ метапоиск
- Примерная стратегия построения ответа:
  – определение типа вопроса
  – построение запроса к интернет-поисковику
  – извлечение из найденных документов нужной информации
  – построение фразы ответа

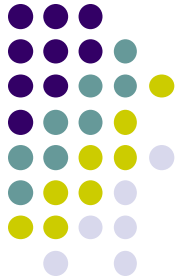
ПРИКЛАДНЫЕ ЗАДАЧИ: *INFORMATION EXTRACTION*



Извлечение информации (знаний) из текстов:

- Специфика задачи – выявление в текстовой коллекции информации, релевантной определенной проблеме, теме:
  - конкретных объектов (имен лиц, названий фирм и т.п.)
  - их отношений , связанных с ними событий и фактов:
    …прошла встреча…, …выдан кредит..
  - терминов и их связей, ключевых слов: *адресная шина*
- Извлеченные данные структурируются и визуализируются
- Приложения:
  - мониторинг новостных лент
    Сколько кораблей затонуло в текущем году?
  - аналитика экономической и производств. деятельности
- Методы извлечения: правила, машинное обучение

ПРИКЛАДНЫЕ ЗАДАЧИ: *OPINION MINING*

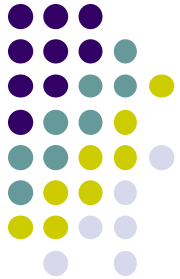


- Близко по целям и методам к направлению *Information Extraction*
- *Opinion Mining* – извлечение и анализ
  - мнений, отзывов, суждений (о персоналиях, товарах, услугах, фильмах, книгах и проч.)
  - из текстов сети Интернет (форумы, блоги и т.п.)
  - их последующей классификации (например, по источнику/ тональности) или др. анализу
- *Sentiment Analysis* – анализ тональности текстов, т.е. определение их общей эмоциональной оценки:

  положительная, *отрицательная*, *нейтральная*

  - о политиках, партиях, фирмах и компаниях и пр. (по сути: задача *контент-анализа)*

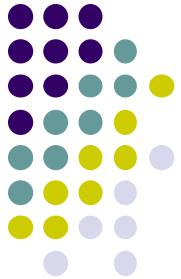
ПРИКЛАДНЫЕ ЗАДАЧИ : *WRITING SUPPORT*



Автоматизация подготовки и редактирования текстов

- Первые программы:
  - автоматическая простановка переносов слов
  - проверка орфографии (спеллеры, автокорректоры)
- Коммерческие системы: проверка орфографии, частично – синтаксиса, а также оценка сложности стиля
- Исследовательские разработки:
  - выявление неправильного употребления предлогов (использование *моделей управления*)
  - обнаружение сложных лексических ошибок: описки, приводящие к другим словам: *овальный/оральный;* паронимические ошибки*: болотный/болотистый*

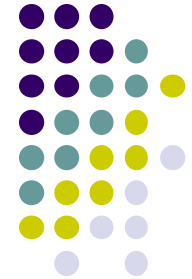
ПРИКЛАДНЫЕ ЗАДАЧИ: ГЕНЕРАЦИЯ ТЕКСТА



С 70-х гг. – в рамках ИИ, рост работ в 90-2000 гг.

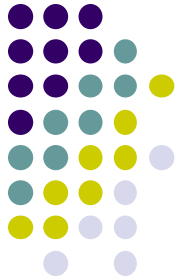
- Особенности задачи:
  автоматическое построение описания на ЕЯ информации, представленной в нетекстовой форме: БД, таблицы, семантические сети, рисунки и др.
  - при этом требуется нужный объем текста и аспект описания
- Виды генерируемых текстов: отчет по БД, комментарий фактов, инструкция пользования
- Примеры систем:
  - Системы многоязыковой генерации (тиражирования) инструкций, руководств пользователя, патентных формул
  - *FoG* (Канада) – двуязычная генерация текстов метеосводок (на английском и французском языках)

ДРУГИЕ ПРИКЛАДНЫЕ ЗАДАЧИ



- Диалог с пользователем на ЕЯ (ИИ, с 60-х гг.)
  - – запросы к специализированной БД
    (язык ограничен лексически и грамматически)
  - – чат-боты (виртуальные собеседники):
    ELIZA (1965 г.), Тест Тьюринга?
    A.L.I.C.E (2000-04), *Rose* (2010-15), *Mitsuku* (2013-16)
  - ❖ разбор вопроса, генерация фразы ответа (по шаблону)
- Обучение ЕЯ:отдельные уровни и модели)
  обычно: лексика языка, грамматика
- Распознавание и синтез звучащей речи:
  - – учет *фонологического* уровня
  - – использование словарей и моделей морфологии

ЗАКЛЮЧЕНИЕ



- Появляются новые прикладные задачи обработки текстов, требующие методов КЛ и анализа данных.
- В большинстве приложений используются простые и редуцированные модели ЕЯ – причина: трудоемкость разработки сложных моделей КЛ, неэффективность применяемых в них алгоритмов.
- Однако простые модели во многих задачах дают приемлемые/хорошие результаты.
- Современная тенденция – все более широкое применение машинного обучения, которое дополняет традиционный *инженерный подход*.

СПАСИБО ЗА ВНИМАНИЕ!

Вопросы?

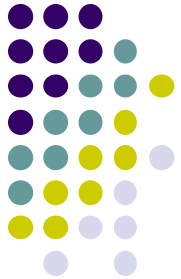
ЛИТЕРАТУРА

- Прикладная и компьютерная лингвистика / Под ред. Николаева И.С. и др. – М.: ЛЕНАНД, **2016**.
- Ингерсолл Г.С., Мортон Т.С., Фэррис Э.Л. Обработка неструктурированных текстов. Поиск, организация и манипулирование / Пер. с англ. – М.: ДМК Пресс, **2015**.
- Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И. и др. – М.: МИЭМ, 2011.

http://clschool.miem.edu.ru/uploads/swfupload/files/98e8cdfb0288b275a3197626ffe06e277a03d43d.pdf

- Васильев В. Г., Кривенко М. П. Методы автоматизированной обработки текстов. – М.: ИПИ РАН, 2008.
- Jurafsky D., Martin J. Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Prentice Hall, 2000.
- Лукашевич Н.В. Тезаурусы в задачах информационного поиска. – М.: Изд-во Московского университета, 2011.

МОДЕЛЬ «СМЫСЛ⇔ТЕКСТ»



И.А. Мельчук, Ю.Д. Апресян (примерно с 70-х гг.)

Лингвистическая структурная модель на основе правил; Особенности:

- *Смысл* – инвариант синонимич. преобразований текста
- Ориентация на синтез (построение) текстов
- Многоуровневость модели, разделение основных уровней на *поверхностный* и *глубинный* уровни, в частности:
  глубинный (семантизированный) и
  поверхностный («чистый») синтаксис
- *Лексические функции* для описания нестандартной синтактики
- Упор на словарь, а не на грамматику,
  в словаре – информация для разных уровней языка
- Семантическое представление предложения/текста:
  семантический граф + коммуникативная организация