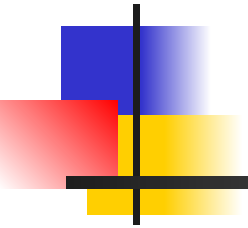
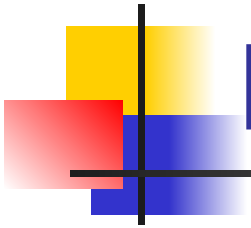


ПРОГРАММНАЯ СИСТЕМА ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ТЕКСТОВ (ПС INEX)

Исследовательский центр искусственного интеллекта
Института программных систем РАН
г. Переславль-Залесский



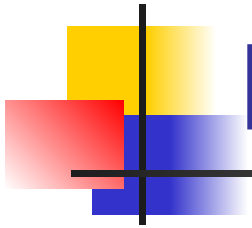
Цели и задачи

Основная цель:

- разработка технологических программных средств извлечения информации из текста

Задачи:

- язык описания правил извлечения информации
- методы предварительной обработки текстов
- среда применения правил извлечения информации
- использование преимуществ параллельной архитектуры



Извлечение информации

Цель:

- извлечь значимую информацию определенного типа из (больших массивов) неструктурированного текста для дальнейшей аналитической обработки

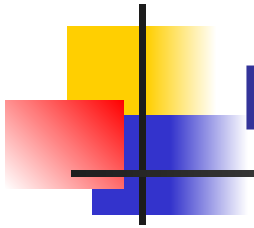
Результат:

- заполненные структуры данных предопределенного формата (экзофреймы)

Примеры предметных областей

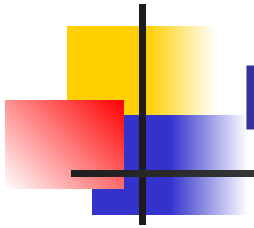
- **Спортивные события:**
  <победитель>, <проигравший>, <счет>, <место_встречи>, <дата>…
- **База данных о рынке жилья:**
  <район>,<цена>, <количество_комнат>, <контактный_телефон>…
- **База данных новых товаров:**
  <производитель>, <дата выпуска>, <название товара> …

Приложения технологии извлечения информации

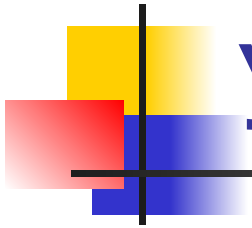


- семантическая кластеризация и классификация
- автоматическое аннотирование
- визуализация данных
- семантическое сравнение и поиск
- создание баз данных
- ...

Извлечение информации: проблемы



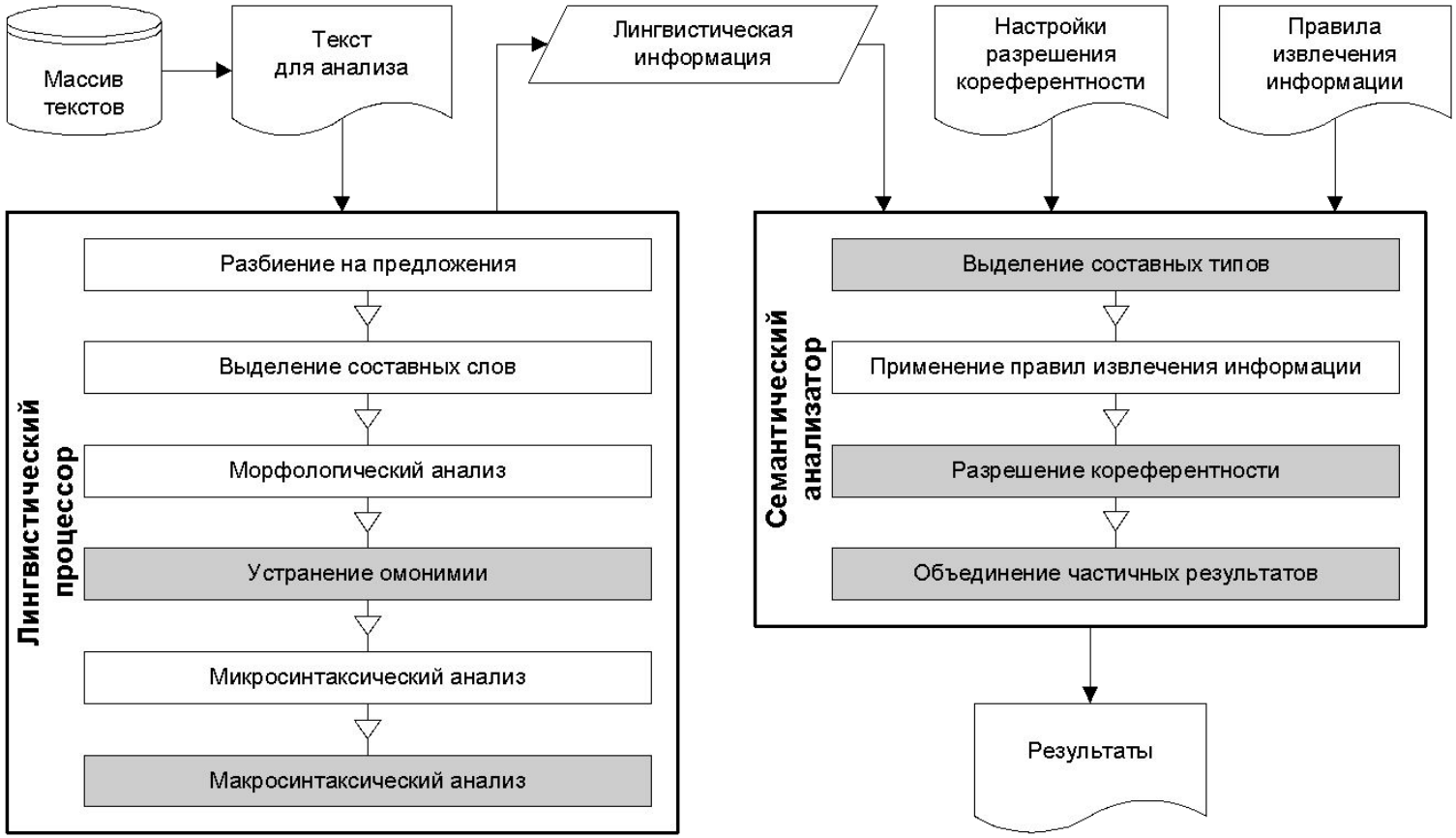
- Необходима точная постановка задачи
- Специфика предметной области
  - лексикон
  - стиль изложения
  - различный характер данных
- Неоднозначности на большинстве этапов обработки текста
- Трудоемкость разработки и настройки систем

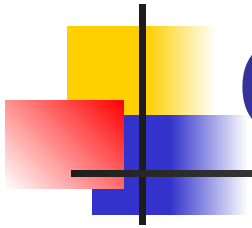


Уровни анализа текста

- графематический анализ
- морфологический анализ
- синтаксический анализ
- прикладной семантический анализ
  - определение семантических классов
  - разрешение кореферентности
  - объединение результатов
  - построение модели предметной области

Архитектура системы извлечения информации

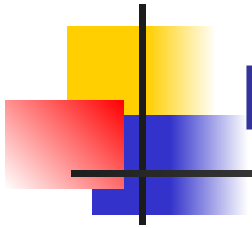




Организация библиотеки

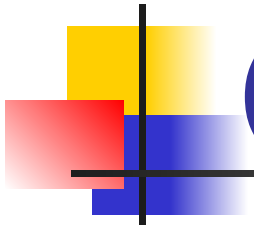
- Документы
- Аннотации
- Итераторы
- Фильтры
- Прикладные задачи
- Анализаторы
- Представления
- Фреймы результатов
- Подсистема ввода-вывода

Подходы к представлению информации о тексте



- Объектные модели ОО-языков
  - высокое быстродействие
  - вероятность сбоев
  - сложность обмена данными и интеграции средств
- Универсальные способы
  - гибкость

Базовая модель аннотаций (TIPSTER)

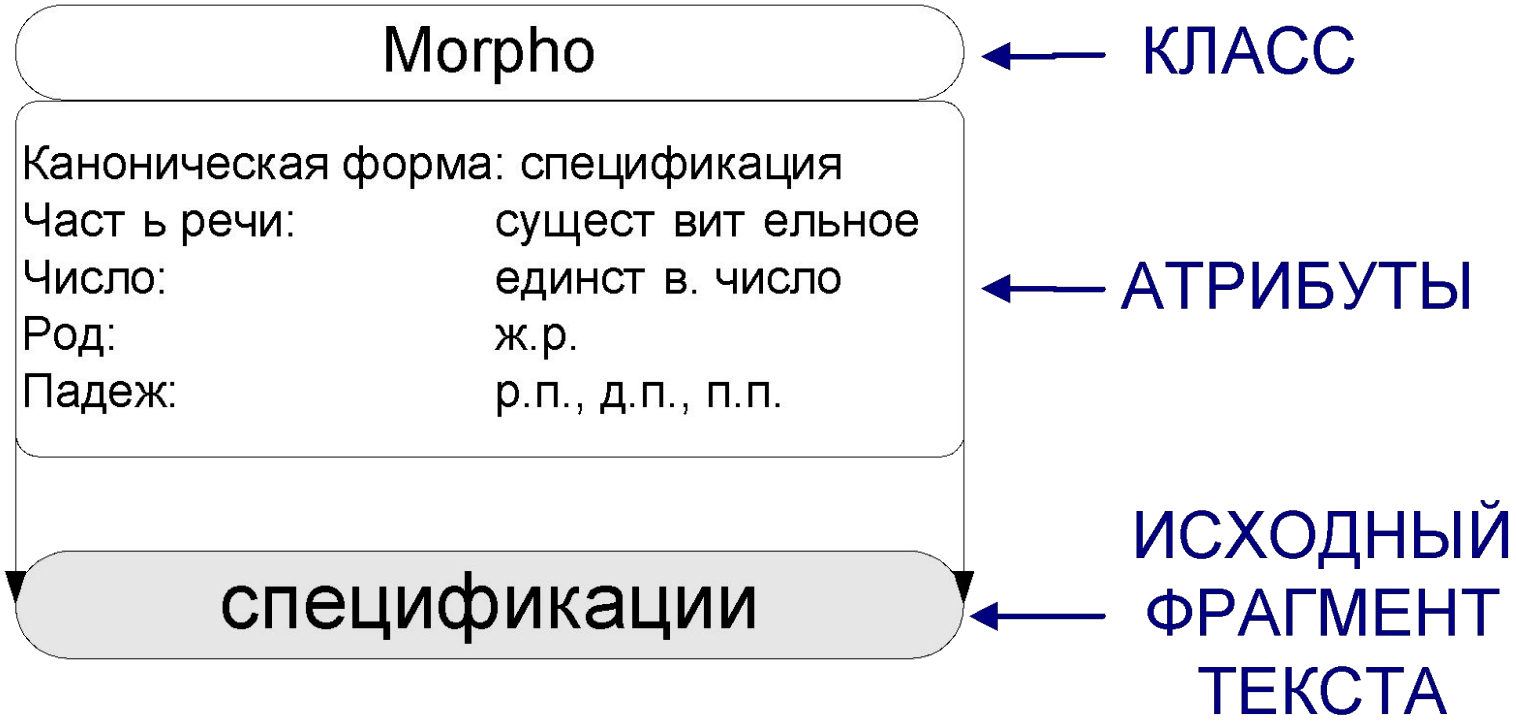


Аннотация

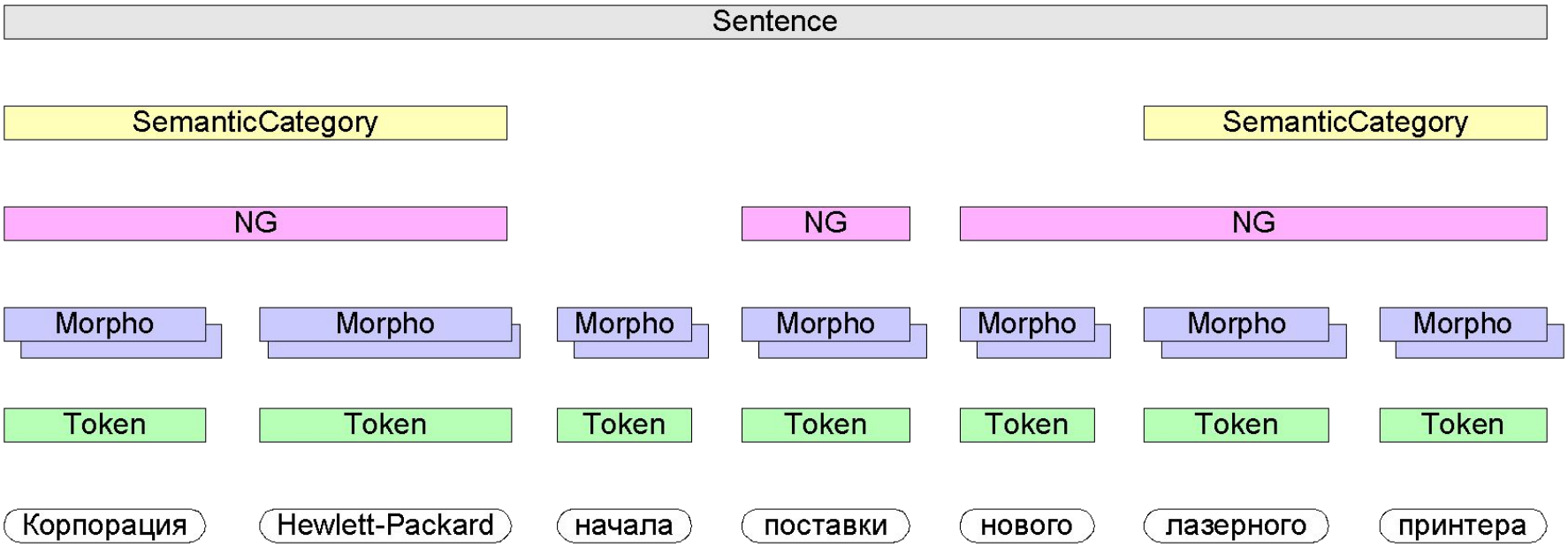
- сопоставляется фрагменту текста;
- принадлежит классу аннотаций;
- содержит атрибуты в виде «имя-значение».

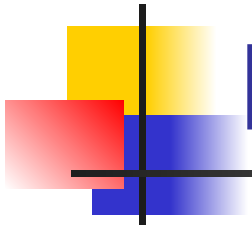
Представление информации о тексте в ПС INEX

ПРИМЕР АННОТАЦИИ



Аннотации: пример





Преимущества

- Унифицированный способ представления информации
- Построение систем со слабой связностью
- Наличие математической модели
- Удобство сопоставления образцу

Пример анализа текста

ФРАГМЕНТ ЛЕНТЫ НОВОСТЕЙ

Японская фирма Victor Company of Japan представила новый DVD-проигрыватель JVC XV-A707 с возможностью воспроизведения дисков DVD-Audio. Подробнее…

Компания MAS Elektronik представила новый стационарный DVD-рекордер Xoro HSD R545 со встроенным ТВ-тюнером и возможностью записи дисков стандарта DVD+R/RW.
Подробнее…

Пример анализа текста

ЦЕЛЕВЫЕ ФРЕЙМЫ

Производитель	*Victor Company of Japan*
Тип	*DVD-проигрыватель*
Модель	*JVC XV-A707*
Носители	*DVD-Audio*

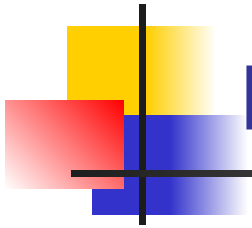
Производитель	*MAS Elektronik*
Тип	*DVD-рекордер*
Модель	*Xoro HSD R545*
Носители	*DVD+R/RW*

Пример анализа текста

ИЗВЛЕЧЕННАЯ ИНФОРМАЦИЯ В СТРУКТУРИРОВАННОМ ВИДЕ

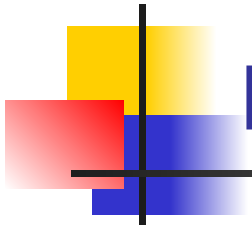
Производитель	Тип	Модель
Victor Company of Japan	DVD-проигрыватель	JVC XV-A707
MAS Elektronik	DVD-рекордер	Xoro HSD R545
Denon	DVD-проигрыватель	AVR-550SD
…		

Правила извлечения информации



- набор правил, описывающих способ извлечения информации и заполнения слотов целевого фрейма
- набор ограничений, накладываемых на текстовые единицы при применении правил

Правила извлечения информации



- Работают на графе аннотаций
- Представляют собой расширение идеи регулярных выражений
- Оперируют аннотациями
- Интерпретируются в соответствии с режимом сопоставления