ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ ИЗ ТЕКСТОВ: ПОРТРЕТ НАПРАВЛЕНИЯ

Большакова Елена Игоревна
МГУ им. М.В. Ломоносова, Факультет ВМиК
bolsh@cs.msu.su

СОДЕРЖАНИЕ

1. Особенности задачи
2. Выделяемые сущности
3. Технология решения: шаблоны
4. Проект ONTOS и система GATE
5. Задача извлечения терминологии
   - Особенности терминов и их употребления
   - Критерии распознавания
   - Шаблоны для извлечения

ОСОБЕННОСТИ ЗАДАЧИ

- *Information Extraction*
- Специфика задачи – распознавание и извлечение из текста определенной значимой информации - объектов и фактов, структуризация извлеченной информации
- Приложения:
  - текстовая аналитика (экономическая, производственная, правоохранительная и др.)
  - построение онтологий и тезаурусов, моделей проблемной области

ВЫДЕЛЯЕМЫЕ СУЩНОСТИ

- Именованные сущности:
  - Имена персоналий
  - Географические названия
  - Названия фирм и организаций
  - Адреса
  - Даты
- Отношения (связи) выделенных сущностей, например: *работать в*

  Смирнов А. работает в ОА «Альфа» с 1998 г.
- связанные с ними события и факты

  получение кредита, слияние компаний…

ТЕХНОЛОГИЯ РЕШЕНИЯ

- Частичный синтаксический анализ :
  неэфффективность и многовариантность синт. разбора
- *Лигвистические шаблоны*, содержащие лексическую, морфологическую и синтаксическую информацию
- Лингвистич. шаблон – описание языковой конструкции, ее *лексического состава* и *грамматических* свойств:
  N “работает” в NP (Noun Phrase)
- Элементы шаблонов:
  - Словоформы, лексемы (возможно, с указанием части речи/морфологических характеристик)
  - Грамматические конструкции: именные и др. группы

ПРОЕКТ ONTOS

АвиКомп, 2000 – 2010 гг.

- Извлечение под управлением онтологии
- Инструментальная система GATE
- Семейство систем OntosMiner - для разных ЕЯ и ПО
- Цели
  - Построение модели ПО
  - Семантическая навигация по тексту
  - Дайджестирование
  - Реферирование: основа реферата - извлеченная информация

СИСТЕМА GATE КАК ИНСТРУМЕНТ

- Набор стандартных программных компонент (лингвистических процессоров) для обработки текста
- Представление лингвистической информации об обрабатываемом тексте в виде набора *аннотаций*, которые хранятся отдельно от текста
- Графическая среда для сборки приложения из компонент

GATE: ПРИМЕРЫ АННОТАЦИЙ

Сущность «Angela Merkel»

Вид аннотации,позиции в тексте			Содержание аннотации
Lookup	41	47	majorType=person_first, minorType=female
Person	41	54	gender=female, rule=PersonFinal, rule1=PersonFull

Token	41	47	category=NNP, kind=word, length=6, orth=upperInitial, string=Angela
Token	48	54	category=NNP, kind=word, length=6, orth=upperInitial, string=Merkel

GATE : КОМПОНЕНТЫ

Цепочка обработки текста в системе GATE:

- **Tokeniser** - разбиение текста на отдельные токены (числа, знаки препинания, слова)
- **Gazetteer** - создание аннотаций к словам на основании словарных файлов (названия городов, организаций, дней недели и т.д.)
- **Sentence Splitter** - разбиение текста на предложения
- **Part of Speech Tagger** - определение части речи слов на основании словаря и правил
- **Semantic Tagger** - распознавание языковых конструкций и сущностей на основе аннотаций и JAPE-правил
- **OrthoMatcher** (Orthographic Coreference) - соотнесение идентичных сущностей с разными названиями

GATE : ШАБЛОНЫ И ПРАВИЛА

Язык **JAPE** - запись правил преобразования аннотаций

- Шаблоны для выявляемых конструкций, например:

  {Morph.SpeechPart="Adjective", Morph.Case="Nominative"}
  - шаблон для выявления прилагательных в именит. падеже

- Правила для преобразования аннотаций :

  левая часть – шаблон, правая – преобразование нужных аннотаций выявленной конструкции

  Rule: Second_name
  ({Token.SemanticType="Name: FName"}):family
  {[А-Я]}{Token.Text="."}{[А-Я]}{Token.Text=="."}) →
  family.Family={rule="Second_name"} -

  правило для выявления имен персоналий вида *Иванов И.* и выделение из них фамилий

ИЗВЛЕЧЕНИЕ ТЕРМИНОВ и СВЯЗЕЙ

- Терминологические слова и словосочетания:
  называют понятия проблемной области:

  общий регистр, *число с плавающей точкой*
  технология двойной накачки
- Приложения:
  - индексирование текстов
  - навигация по тексту
  - поддержка терминологич. редактирования текстов
  - построение *глоссариев* и *предметных указателей*
  - создание онтологий и тезаурусов

  Часть приложений – обработка отдельного текста, но не коллекции

ОСОБЕННОСТИ ТЕРМИНОВ

- Большинство словосочетаний – *несвободные* (некомпозиционные), т.е. их смысл не выводится из смысла компонент:

  железная дорога, *длина слова*

- Конвенциональность научно-технических терминов ⇒ необходимость их определения в тексте:

  Под прерыванием понимается сигнал…

- Грамматическая структура терминов: чаще всего - именные словосочетания, их можно описать структурными грамматическими образцами:
  - прилагательное-существительное – *логический вывод,*
  - существительное- существительное в род. падеже – *период упреждения*

МЕТОДЫ РАСПОЗНАВАНИЯ

- Применение статистических и лингвистических критериев:
- Статистические критерии

Например, функция упорядочивания по статистике:

$$\text{C-Value}(\mathbf{a}) = \begin{cases} \log_2 |\mathbf{a}| * \text{freq}(\mathbf{a}), \text{ если } \mathbf{a} \text{ не вложено, иначе} \\ \log_2 |\mathbf{a}| * (\text{freq}(\mathbf{a}) - P(T\mathbf{a})^{-1} * \sum_{b \in T\mathbf{a}} \text{freq}(b)) \end{cases}$$

- где **a** – слово (словосочетание), $|\mathbf{a}|$ – его длина,

freq(**a**) – частота встречаемости **a** в тексте,

T**a** – множество словосочетаний текста, содержащих **a**,

P(T**a**) – количество словосочетаний, содержащих **a**.

электрический слой - двойной электрический слой

МЕТОДЫ РАСПОЗНАВАНИЯ: ЛИНГВИСТИЧЕСКИЕ КРИТЕРИИ

- грамматические (синтаксические) образцы терминов:

  A N N - *спектральный коэффициент излучения*

- контексты употребления терминов:

  effect of T – effect of drought, effect of cold
  (последствие засухи, заморозков)

  such T1 as T2 – such crimes as money laundering
  (такие преступления, как отмывание денег)

- Лингвистическую информацию можно записать в виде шаблонов

  необходим язык шаблонов и поддерживающие его средства

РАСПОЗНАВАНИЕ ТЕРМИНОВ: ТЕКСТОВЫЕ ВАРИАНТЫ

При использовании терминов в тексте они могут образовывать варианты:

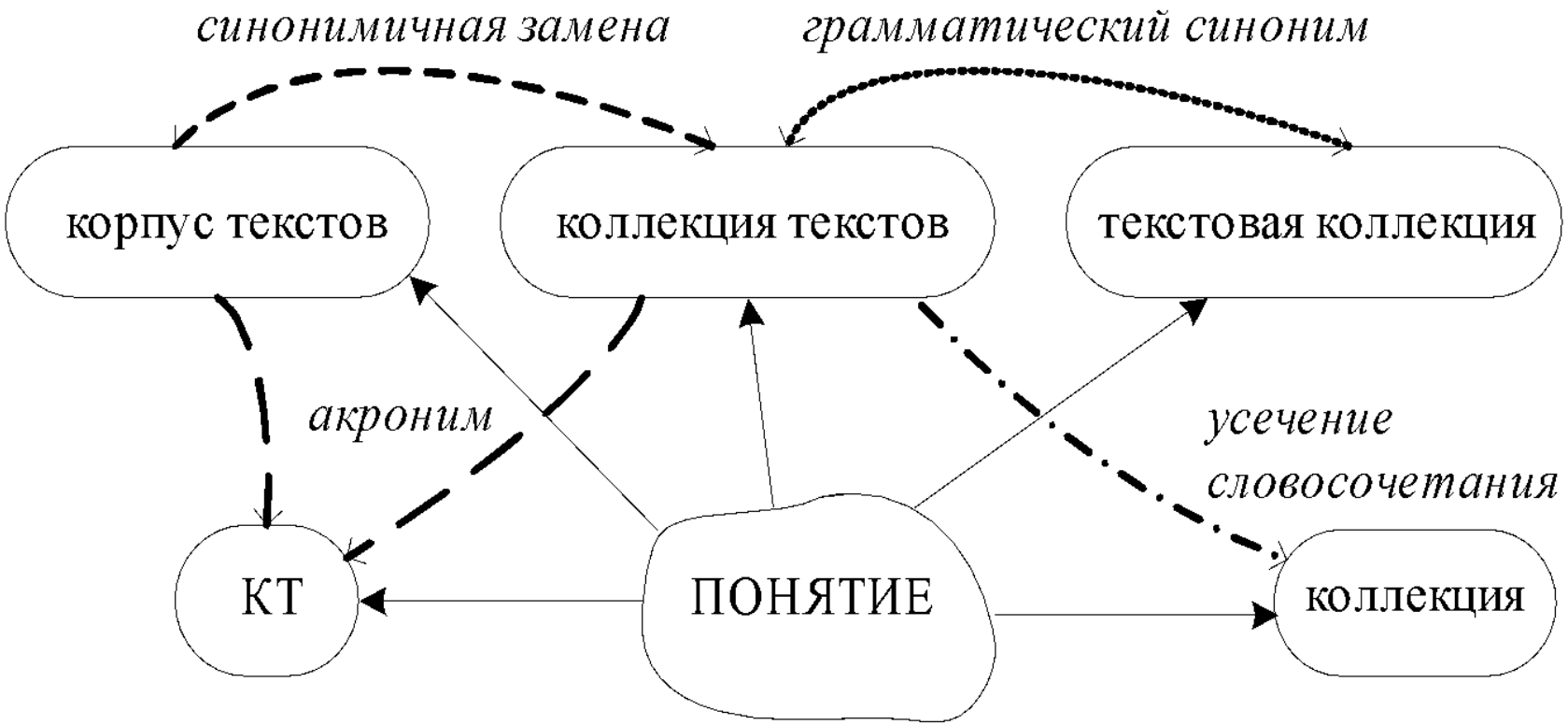
- Орфографические варианты: *браузер - броузер*
- Морфоварианты: *спецсимвол – спецзнак*
- Лексико-синтаксические варианты:

  механическое напряжение - напряжение

  дисковый контроллер – контроллер диска
- Варианты сокращений: *ЦП, авост*

В словаре представлены далеко не все варианты терминов, их необходимо распознвать

ТЕРМИНОЛОГИЧЕСКИЕ ВАРИАНТЫ: ПРИМЕР



РАСПОЗНАВАНИЕ ТЕРМИНОВ: СОЕДИНЕНИЯ ТЕРМИНОВ

Соединения нескольких терминологических словосочетаний:

- Бессоюзные соединения, с разрывом и без разрыва термина:

  разрядность внутренних регистров

  – *разрядность регистра*, *внутренний регистр*

- Соединения с союзом:

  шинам адреса, данных и управления

  – *шина адреса*, *шина данных*, *шина управления*

Средство распознавания - лингвистические шаблоны

ШАБЛОНЫ: ЯЗЫК LSPL

Лексико-синтаксический шаблон позволяет задать для элемента-слова:

- часть речи (A, N, V, Pa и т.д.) – A
- индекс – A1 A2 N
- лексему – A<важный>
- морфологические характеристики (имя=значение) – A<важный; case=nom, gen=fem>

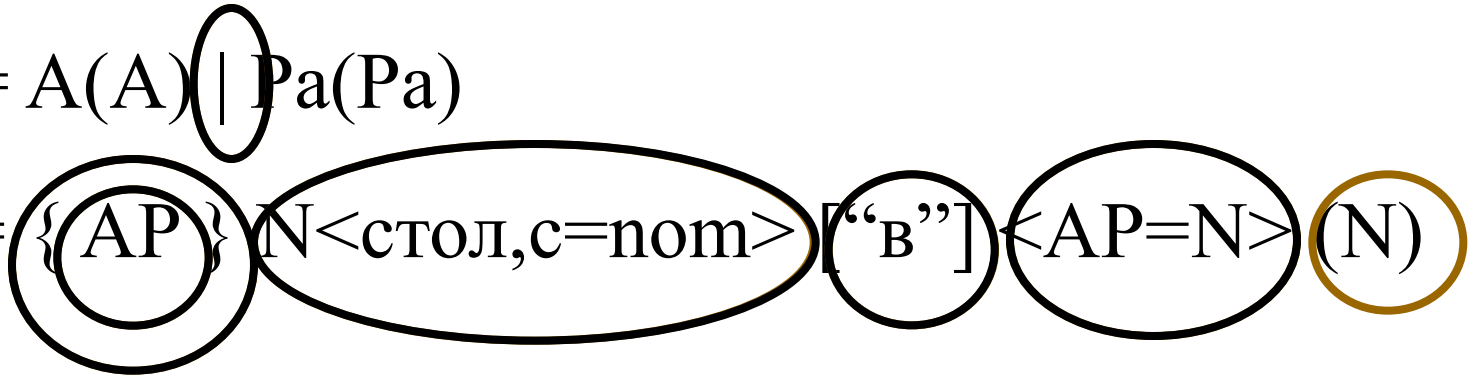
Грамматическое согласование элементов шаблона:

A<тяжелый> N <A.gen=N.gen, A.num=N.num, A.case=N.case>

Прилагательное *тяжелый* и существительное согласованы в роде, числе и падеже: *тяжелым вечером*, *тяжелых камней*, *тяжелое тело*

ЯЗЫК LSPL-ШАБЛОНОВ: ВОЗМОЖНОСТИ

- AP = A(A) | Pa(Pa)
- AN= {AP} N<стол,с=nom> [“в”] <AP=N> (N)



Элемент-слово
Имя шаблона
Экземпляр шаблона
Условия согласования

Альтернативы |
Повторение {}
Опциональное вхождение []
Параметры шаблона

LSPL-ШАБЛОНЫ: ПРИМЕРЫ

- Шаблон типичной структуры термина:

  A N1 { N2 <case=gen>} (A=N1)

  реактивная сила, немаркированный квантор общности

- Шаблон типичной фразы-определения новых терминов:

  NP1<c=acc> ["мы"] "назовем" NP2<c=ins> <NP1.n = NP2.n>

  *Указанную операцию **назовем** операцией поиска примеров*

- Шаблон образования терминологических вариантов:

  N1 N2<c=gen> "," N3<c=gen> {"и"|"или"} N4<c=gen>

  #N1 N2<c=gen> , N1 N3<c=gen> , N1 N4<c=gen>

  шинам адреса, данных и управления –

  шина адреса, шина данных, шина управления

ЗАКЛЮЧЕНИЕ

- В основном – извлечение на основе правил (*rule-based*), все чаще - машинное обучение
- Точность и полнота извлечения
  - зависят от набора шаблонов
  - зависят друг от друга
  - верхняя граница - до 80-90 %
- Сложность задачи (технологическая): приемлемая полнота и точность достигается
  - на больших массивах текстов
  - обычно в рамках коммерческих компаний

СПАСИБО ЗА ВНИМАНИЕ!