ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ ИЗ ТЕКСТОВ (*INFORMATION EXTRACTION*)

ОСОБЕННОСТИ ЗАДАЧИ

- ***Information Extraction*** *– перевод неструктурированной информации в структурированную*.
- Специфика:
  - обработка текста или коллекции текстов
  - экстрагирование семантически значимых данных, релевантных (по определенной проблеме, теме, вопросу)
  - структуризация извлеченных данных (таблицы, шаблоны),
  - накопление информационной базы, визуализация данных
- Например, поиск информации о деловых визитах:
  1 апреля 2012 г. представители компании Яндекс посетили офис госкорпорации «Нанотехнологии»…
- Приложения:
  - мониторинг новостных лент (сколько кораблей затонуло в текущем году?)
  - Составление дайджестов и рефератов
  - аналитика деятельности: экономической, производственной, правоохранительной и др.

ИЗВЛЕКАЕМАЯ ИНФОРМАЦИЯ

- Именованные (конкретные) сущности

  NE - Named Entities

  - персоны, компании, адреса, даты
  - упоминания генов и белков и пр.
- Отношения выделенных сущностей:
  - Место работы, должность
  - Взаимодействие белков
- Связанные с ними события и факты

  Events

  слияние/поглощение компаний…
  приобретение контрольного пакета акций

ИМЕНОВАННЫЕ СУЩНОСТИ

- Имена персоналий: *И.Сечин, Ben White, Ng A.*
- Географические названия: *Apple Str.*
- Названия фирм, компаний, организаций
- Адреса
- Даты и временные отрезки
- Марки товаров
- Ссылки на литературу
- Гены, белки, хим. вещества,

Особенности:

- Большое количество разных имен
- Постоянно появляются новые сущности
- Нет строгих правил именования: *«Путь к себе»*

ОТНОШЕНИЯ СУЩНОСТЕЙ

- Атрибуты конкретных объектов:
  - Должность: *инженер, менеджер*
  - Телефон
  - Место работы
  - Цвет, размер
- Отношения (связи) конкретных объектов,

  например: *работать в*

  быть владельцем
- Виды отношений:
  - общие: причина-следствие, цель-результат
  - зависящие от тематики, предметной области текста

СОБЫТИЯ - ***EVENTS***

- Пример:
  Очередной долгосрочный государственный кредит в размере 300 млн. евро получило правительство Греции
- Извлекаются:
  - имена сущностей
  - их атрибуты, отношения
  - факты
- Структурирование:

фрейм получения кредита:

атрибуты:	сумма	получатель	вид
	300 млн.	пр-во Греции	долгосрочный

ПОДХОДЫ К ИЗВЛЕЧЕНИЮ

- Основанный на машинном обучении
  - Опора на статистические (вероятностные методы)
  - Необходим размеченный вручную текстовый корпус
- Основанный на правилах, или инженерный
  rule-based, knowledge-based
  - Извлечение на основе лингвистических правил
  - Правила извлечения пишутся экспертами
  - Используются специальные языки записи правил и поддерживающие их программные инструменты
- Современные тенденции: комбинирование
- Общее – очень ограниченный набор тем извлекаемой информации

МАШИННОЕ ОБУЧЕНИЕ ДЛЯ *IE*

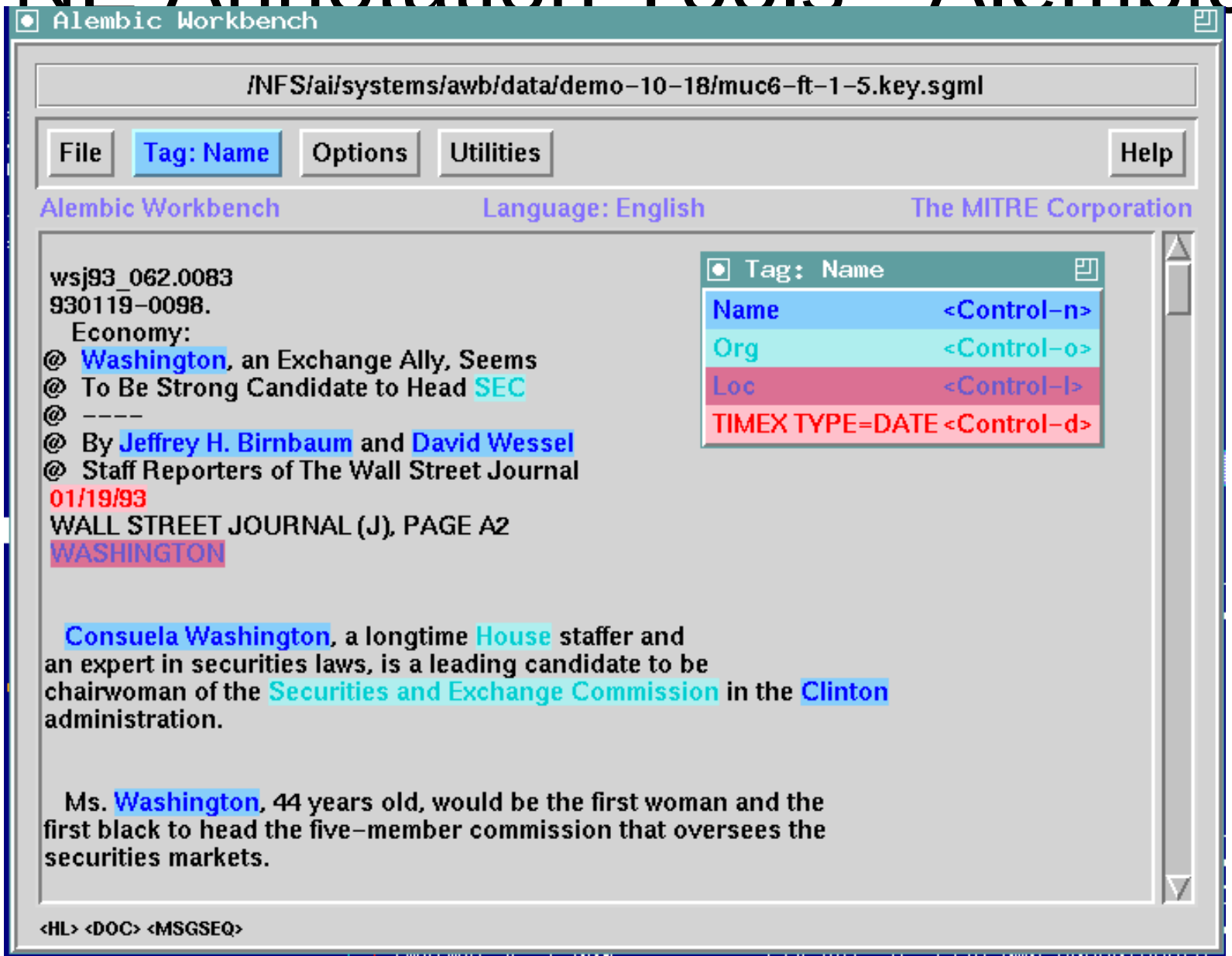
Дано: коллекция текста, в котором отмечены выделяемые объекты.

Для построения модуля тегирования объектов используются:

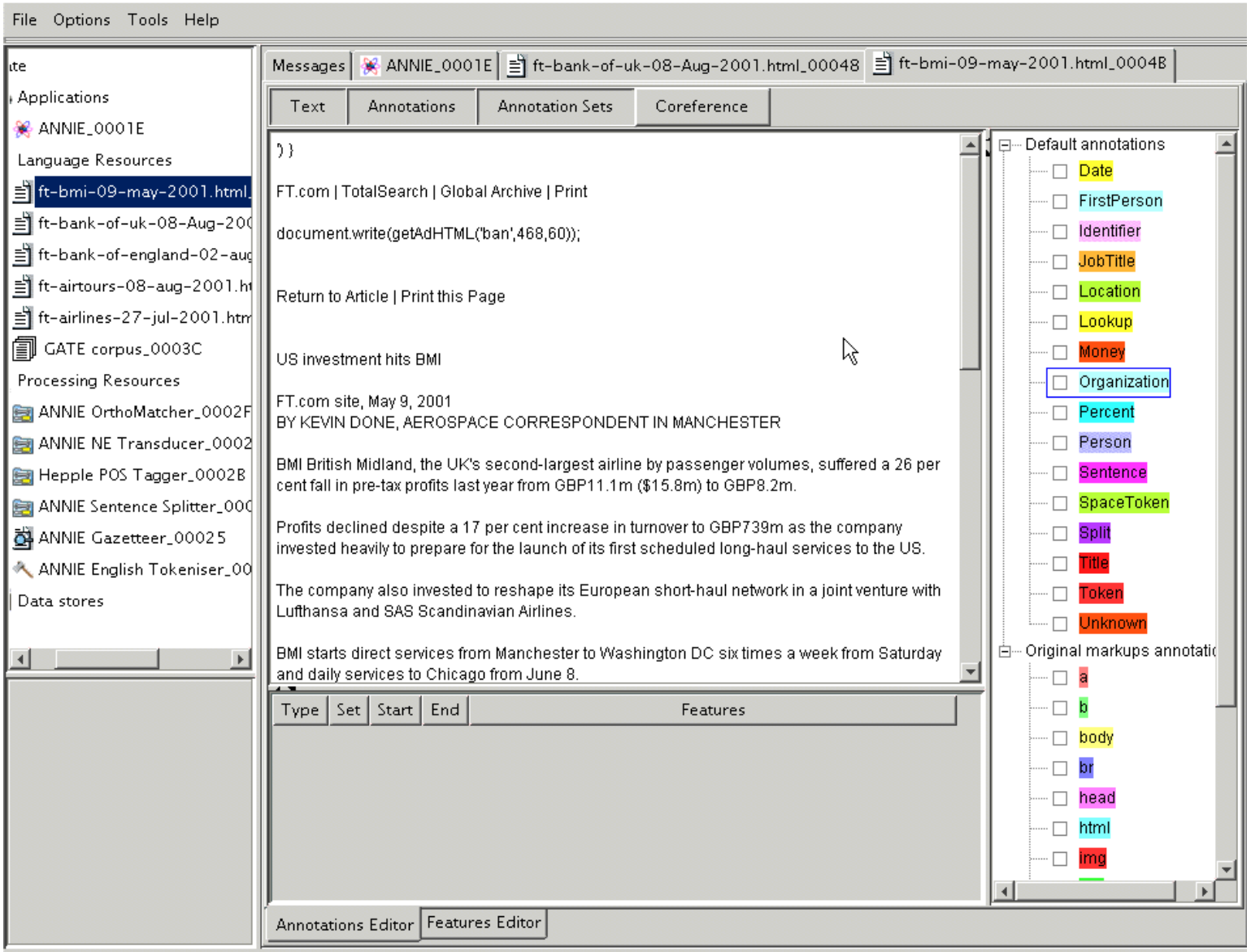
- Простые вероятностные модели
- Деревья решений – *DT*
- Модель максимальной энтропии – *ME*
- Скрытая марковская модель – *HMM*

  – Обучение хорошо работает для извлечения именованных сущностей
  – используется методология *bootstrapping*

NE Annotation Tools - Alembic



NE Annotation Tools - GATE



ИЗВЛЕЧЕНИЕ НА ОСНОВЕ ПРАВИЛ

Особенности инженерного подхода:

- Частичный, поверхностный синтаксический анализ текстов – *shallow approach*, *shallow parsing*

  причина: неэфффективность и многовариантность полного синтаксического разбора
- Лингвистические правила – *лингвистические шаблоны*, содержащие лексическую, морфологическую и синтаксическую информацию об извлекаемой единице текста
- Написание правил экспертами – трудозатратный, небыстрый процесс (+ отладка правил)

ОСНОВНЫЕ ЭТАПЫ ОБРАБОТКИ

- Графематика (токенизация)
  - Разбиение сложных слов (?)
- Морфологический анализ
  - Определение части речи
  - Определение грамматических характеристик
- Лексический анализ
  - Сопоставление со словарями
  - Разрешение лексической многозначности
- Синтаксический анализ
  - Частичный анализ $\Rightarrow$ *лингвистические шаблоны*

    (на базе конечных автоматов, регулярных и контекстно-свободных грамматик)
- Дискурсивный и предметный анализ
  - Анализ референциальных ссылок
  - Слияние (объединение) извлеченных фактов

ЛИНГВИСТИЧЕСКИЕ ШАБЛОНЫ

- Лингвистический шаблон – описание языковой конструкции, ее лексического состава и грамматических свойств:

  N “работает” в NP *N – существительное*

  N “consists of” N *NP (*Noun Phrase) – группа существительного

- Элементы шаблонов:
  - Словоформы, лексемы (возможно, с указанием части речи/морфологических характеристик)
  - Грамматические образцы: именные и др. группы

  A + N + Ngen – спектральный коэффициент излучения
- Шаблоны предполагают частичный синт. анализ
- Для удобства описания шаблонов и правил созданы специальные языки: JAPE, RCO, LSPL и поддерживающие их программные средства

ОСОБЕННОСТИ ИЗВЛЕЧЕНИЯ ИМЕН

- Словарь имен
- Словарь частей имен
- Правила и шаблоны:
  - слова, написанные в определенном регистре (с большой буквы или всеми большими буквами)
  - Слова с определенными последовательностями букв (например: *-ов, -дзе* и др. для фамилий; *-банк, -инвест* для коммерческих организаций)
  - Использование внутренней структуры (ООО)
  - Проверка по корпусу
    - Michigan State – название университета,
    - New York State – название штата
- В итоге: список правил

СЛОЖНОСТИ ИЗВЛЕЧЕНИЯ

- Текст: множество разных именований одной и той же сущности:
  - *William H. Gates, Mr. Gates, Bill Gates*
  - Местоимения: *It/she/he* vs. *он/она/оно*
  - Сокращения
  - Именные группы (*владелец Microsoft*)

  т.е. выделение словесных оборотов, ссылающихся на один и тот же объект
- Общая проблема разрешения кореференциии и анафоры (*Coreference Resolution)*
- При извлечении событий и фактов: необходимость слияния (*Merging*) частичных описаний, полученных из разных предложений
  - необходимы специальные правила, основанные на сопоставлении слотов фрейма,идентификация по связям и по контексту, энциклопедические знания

ОЦЕНКА СИСТЕМ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ

Конференция **MUC** - Message Understanding Conference (1987-1997)

- MUC-1 (87), MUC-2 (89) – Военно-морские операции
- MUC-3 (91), MUC-4 (92) – Террористическая деятельность в Южной Америке
- MUC-5 (93) – Совместные предприятия в области микроэлектроники
- MUC-6 (95) – Служебные перемещения: назначения и отставки
- MUC-7 (97) – Отчеты о запусках космических кораблей и ракет

КОНФЕРЕНЦИЯ MUC-7

Запуски космических кораблей и ракет

Атрибуты события:

- Запущенный_аппарат
- Боевая_часть
- Дата_запуска
- Место_запуска
- Тип_задания (военный, гражданский)
- Цель_запуска (тестирование, доставка..)
- Статус_запуска (удачный, неудачный, выполняется, планируется)

МЕРЫ ЭФФЕКТИВНОСТИ ИЗВЛЕЧЕНИЯ

- Точность (Precision)

$$\boldsymbol{P} = \frac{\text{к-во правильных ответов}}{\text{к-во полученных ответов}}$$

- Полнота (Recall)

$$\boldsymbol{R} = \frac{\text{к-во правильных ответов}}{\text{общее к-во правильных ответов}}$$

- Соотношение между точностью и полнотой

$$\boldsymbol{F\text{-}Measure} = \frac{(\beta^2 + 1)PR}{\beta^2 R + P}, \text{ где}$$

β – коэффициент относительной важности, обычно β=1

ОЦЕНКИ ЭФФЕКТИВНОСТИ

- Извлечение именованных сущностей:
  (*Wall Street Journal)*
  - Машинное обучение (HMM)
    MUC-6 – F=93, MUC-7 – F=90.4
  - Извлечение на основе правил:
    MUC-6 – F=96.4, MUC-7 – F=93.7
- Извлечение отношений
  - На основе обучения – 60-70%
  - На основе правил – высокая точность, низкая полнота
- Извлечение событий, фактов
  - На основе обучения- 50-60%
  - На основе правил: 90% точности, 20% полноты

МАШИННОЕ ОБУЧЕНИЕ ДЛЯ ИЗВЛЕЧЕНИЯ ИМЕН

Сколько необходимо данных для обучения:

- BBN
  - 30000 слов – F 81
  - 1.2 млн.слов – F 91
- MITRE
  - 250K слов – F 79
  - 750K слов – F 86
  - 1.2 млн слов – F 87
- 1.2.млн. слов – 1800 газетных статей
- Последовательность разметки тоже важна
- Linguistic Data Consortium – источник данных

ВЫБОР ПОДХОДА К ИЗВЛЕЧЕНИЮ

- Основанный на правилах (инженерный)
  - Имеются словари, списки слов
  - Имеются инженеры по знаниям
  - Мало размеченных данных
  - Нужно максимально возможное качество извлечения (высокая точность, низкая полнота?)
- Использование машинного обучения
  - Нет словарных ресурсов
  - Нет инженеров по знаниям
  - Требуется быстрое построение приложения
  - Размеченных данных много и получение их дешево
  - Достаточно иметь приемлемое качество извлечения

СОВРЕМЕННЫЕ ТЕНДЕНЦИИ

- Комбинирование подходов к извлечению, итеративное построение (***bootstrapping***)
- Подход с машинным обучением:
  - Обучение начинается с небольшого количества размеченных данных,
  - итеративное расширение обучающего множества
  - поиск естественно размеченных данных
- Подход, основанный на знаниях:
  Автоматизация построения шаблонов - для повышения полноты распознавания отношений и фактов
  - Имеется множество сущностей с известными отношениями, например, штаб-квартиры компаний
  - В текстовом корпусе находятся предложения, в которых упоминаются эти пары сущностей.
  - Формируются наиболее вероятные шаблоны, которые проверяются на других текстах и расширяются

ИНСТРУМЕНТЫ: СИСТЕМА GATE

Широко-известная платформа/среда для построения IE-приложений.

Архитектура:

- Набор стандартных программных компонент-*ресурсов* (лингв. процессоров, словарей) для обработки текста
- Внутреннее представление лингв. информации об обрабатываемом тексте в виде набора *аннотаций*, хранимых 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 в ходе обработки текста

- Шаблоны для выявляемых конструкций: например:

  {Morph.SpeechPart="Adjective", Morph.Case="Nominative"}

  - шаблон прилагательных в именительном падеже

- Правила для обработки аннотаций :

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

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

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

ДРУГИЕ ИНСТРУМЕНТЫ IE

Для построения систем обработки русскоязычных текстов:

- Инструментальная система **RCO Pattern Extractor**
  - Архитектура аналогична системе GATE
  - Внутренний язык подобен Jape + регулярные выражения
  - Добавлены прогр. модули обработки русских словоформ

- Язык **LSPL** и поддерживающие его библиотека и среда (Большакова и др.)

  Основное понятие – *лексико-синтаксический шаблон* конструкции, который позволяет задать для слова:
  - часть речи (A, N, V, Pa и т.д.), индекс слова: *A1 A2 N*
  - лексему: *A<важный>*, словоформу, строку: *“важным”,*
  - морфологические характеристики (имя=значение): *A<важный; case=nom, gen=fem>*

  Возможно также:
  - Грамматическое согласование элементов шаблона: *A<тяжелый> N <A=N>* - прилагательное *тяжелый* и существит-ое согласованы в роде, числе и падеже: *тяжелым вечером*
  - Определение вспомогательных шаблонов, которые можно параметризовать и использовать в других шаблонах

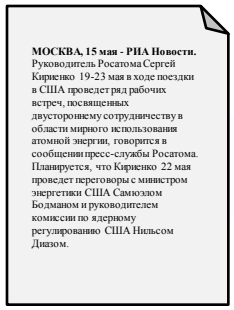
ПРОЕКТ ONTOS

АвиКомп, 2000 – 2012 гг.

- Основа - инструментальная система GATE
- Инженерный подход: извлечение под управлением предметной онтологии, построение ресурсов-обработчиков на основе системы онтологий
- *shallow* подход к анализу текста: *правила анализа изначально ориентированы на распознавание только тех объектов и отношений, которые описываются предметной онтологией.*
- Семейство систем **OntosMiner** - для разных ЕЯ и ПО
- Цели:
  - построение модели ПО текстов
  - генерация дайджестов и рефератов на основе извлеченной информации , в том числе - резюме персоны
  - семантическая навигация по тексту

СХЕМА РАБОТЫ OntosMiner

Текст **Модель** **Структурированные данные**



МОСКВА, 15 мая - РИА Новости. Руководитель Росатома Сергей Кириенко 19-23 мая в ходе поездки в США проведет ряд рабочих встреч, посвященных двустороннему сотрудничеству в области мирного использования атомной энергии, говорится в сообщении пресс-службы Росатома. Планируется, что Кириенко 22 мая проведет переговоры с министром энергетики США Самюэлом Бодманом и руководителем комиссии по ядерному регулированию США Нильсом Диазом.

типы объектов и
типы отношений

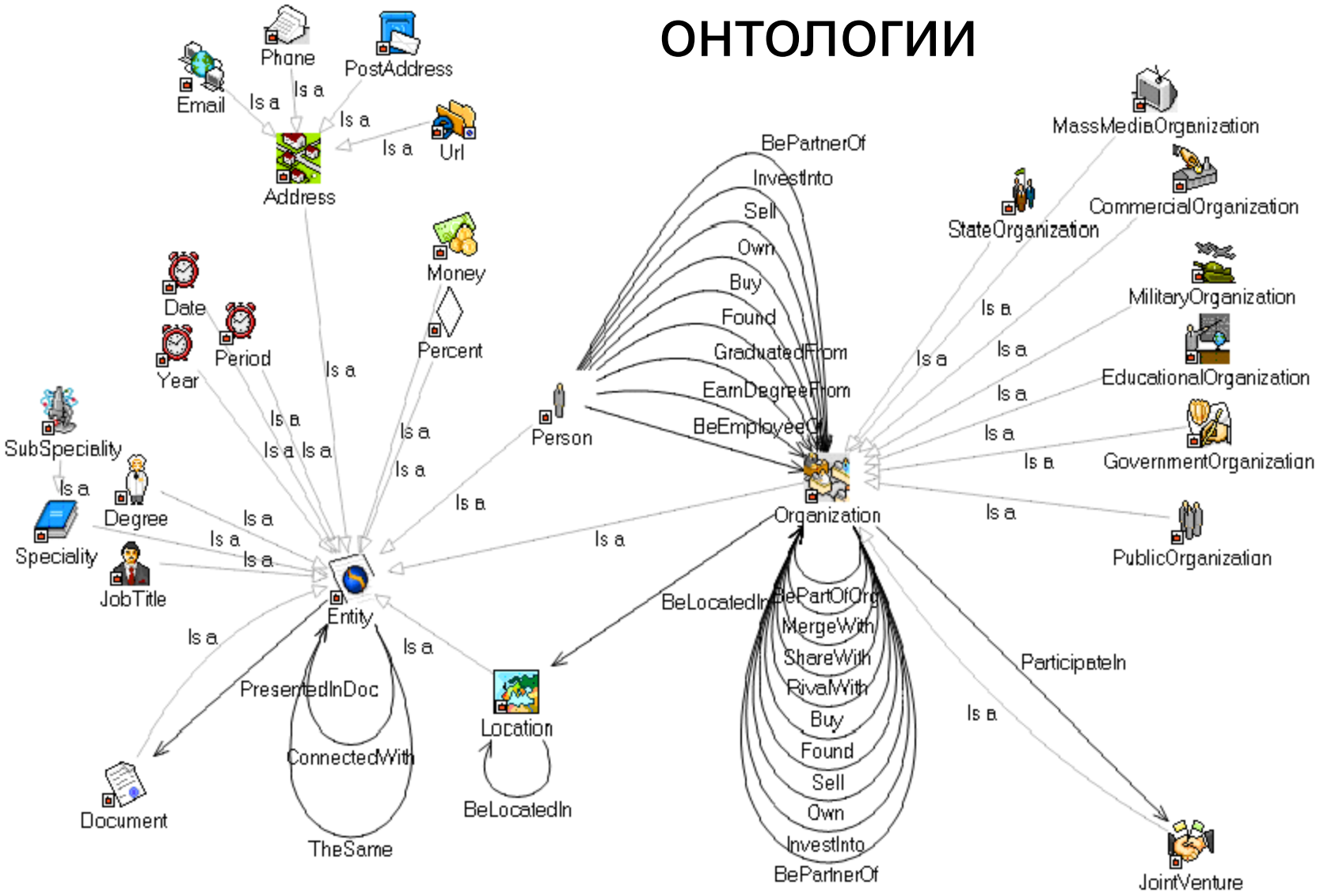
РАБОТАТЬ В ОРГАНИЗАЦИИ

ОРГАНИЗАЦИЯ

ПЕРСОНА

Руководитель Росатома Сергей Кириенко 19-23 мая в ходе поездки в США проведет ряд рабочих встреч…

Пример онтологии



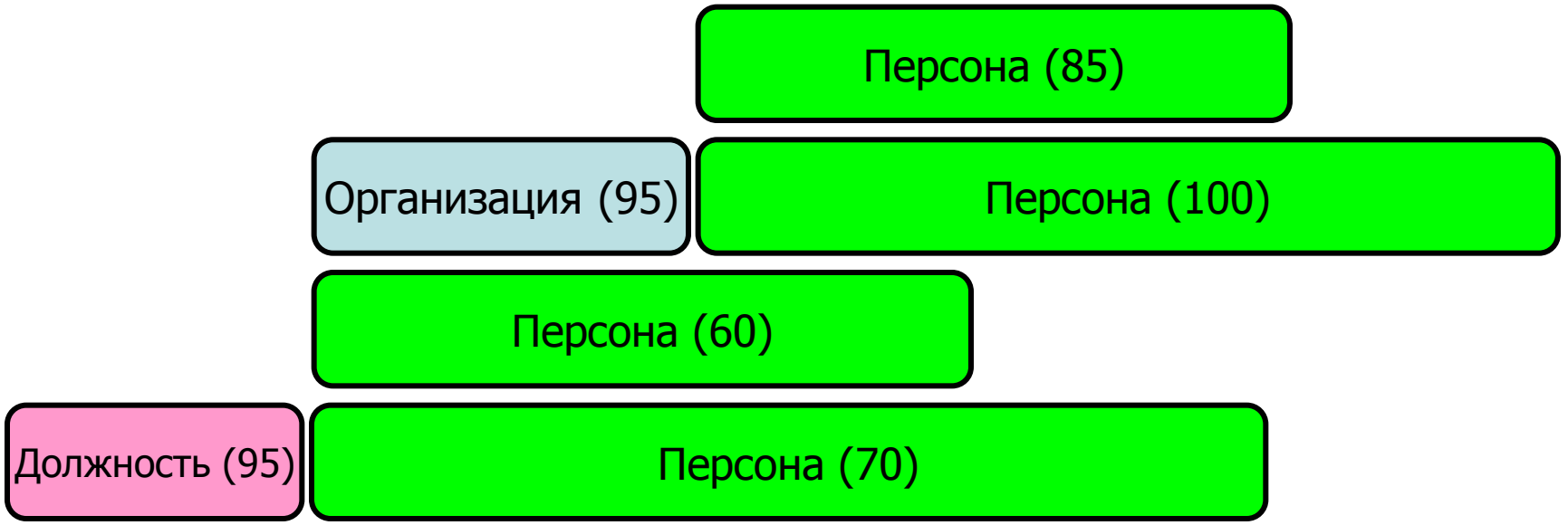
АРХИТЕКТУРА OntosMiner

- Препроцессинг
  - токенизация
  - деление текста на предложения
  - морфологический анализ
  - газетер (*распознавание слов и словосочетаний, релевантных для заданной предметной области*)
- Собственно обработка (на языке Japе)
  - выделение имен (объектов)
  - выделение глагольных групп (отношений)
- Семантическая интерпретация выделенных объектов и групп
  - Обработка синонимов объектов
  - Генерация XML-представления *когнитивных карт*

OntosMiner : ПРИМЕРЫ ПРАВИЛ

Шаблон на языке Jape	Примеры соответствующих фрагментов текста
{AdjNPupper} {Lookup.majorType ==“edu", Lookup.NMB == "sg"} {GenNP} ({GenNP})?	Он поступил в **Московский университет дружбы народов** и отучился там 4 года.
{ JobTitle} {Organization} {Upper}	Председатель совета директоров Газпрома **Миллер** не собирается уходить в отставку

OntosMiner:
Пример неоднозначного сопоставления



Президент Росторгбанка Владимир Семенович Нистратов

- конфликты между выделенными по разным правилам объектами: пересечение, вложение, совпадение
- Способ решения: эвристические веса

Примеры в различных предметных областях

По материалам конференций

Processing news texts

- Very traditional IE boosted by Message Understanding Conferences (MUC) in late 1980s and 1990s (DARPA), followed by Automatic Content Extraction (ACE) initiative and Text Analysis Competition (TAC) (NIST)
- Tasks:
  - Named entity recognition
  - Noun phrase coreference resolution
  - Entity relation recognition
  - Event recognition (who, what, where, when)

Nastia Liukin wins women's gymnastics all-around gold

Adjust font size:

Nastia Liukin of the United States edged her compatriot Shawn Johnson to win the women's all-around after a breathtaking Olympic gymnastics competition on Friday.

WHO?

Finishing one-two on the podium, the American duo let the hosts' gold rush in the gymnastics pause after China wrapped up all the first three titles (men's team, women's team, men's all-round) in the National Indoor Stadium.

WHEN?

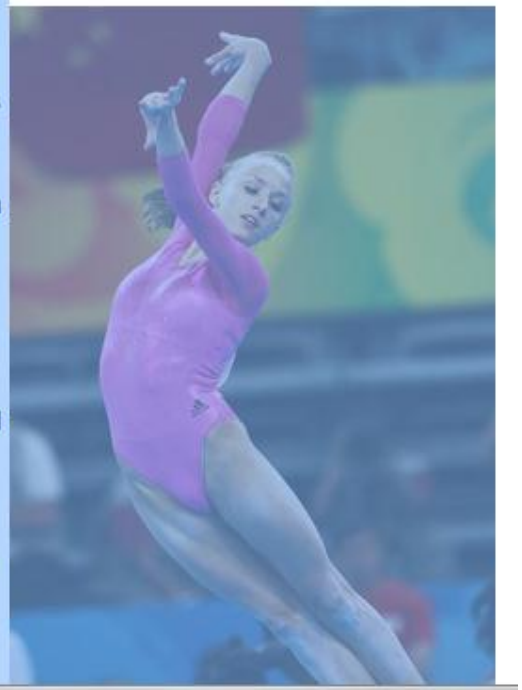
Liukin collected 63.325 points, after flawless exercises on each of the four apparatus with a good combination of difficulty and quality, beating Johnson by 0.600. The bronze medal went to Yang Yilin of China in 62.650.

WHAT?

Coming out in her first Olympic Games, the 18-year-old Liukin struck the most coveted gold medal which she had waited for years.

WHERE?

She was an unlucky runner-up in the 2005 world championship in Melbourne, beaten by Chellsie Memmel by 0.001 points in her debut to the international arena, and the title also eluded her in the following two championships.



www.china.org.cn

Processing news texts

- **Named entity recognition**:
  - Person, location, organization names
  - Mostly supervised: Maxent, HMM, CRF
  - Approaches human performance: in literature sometimes above 95% F_1 measure

[Bikel et al. ML1999] [Finkel et al. 2006]

- **Noun phrase coreference resolution**:
  - Although unsupervised (clustering), and semi-supervised (co-training), best results with supervised learning: F_1 measures of 70% and more are difficult to reach; also kernel methods

[Ng & Cardie ACL 2002] [Ng & Cardie HLT 2003] [Versley et al. COLING 2008]

Processing news texts

- **Entity relation recognition**:
  - use of supervised methods: e.g., kernel methods: F_1 measures fluctuate dependent on number of training examples and difficulty of the relational class (ambiguity of the features)

[Culotta & Sorensen ACL 2004] [Girju et al. CSL 2005]

- **Event recognition**:
  - in addition: recognition and resolution of:
    - temporal expressions: TimeML
    - spatial expressions: FrameNet and Propbank

[Pustejovsky et al. IWCS-5 2003] [Baker et al. COLING-ACL 1998]
[Morarescu IJCAI 2007] [Palmer et al. CL 2005]

Processing biomedical texts

- Many ontologies or classification schemes and annotated databases are available:
    - E.g., Kyoto Encyclopedia of Genes and Genomes, Gene Ontology, GENIA dataset
- Tasks:
  - Named entity recognition
  - Relation recognition
  - Location detection and resolution

A UNIQUE CONTRIBUTION OF HEAT SHOCK TRANSCIPTION FACTOR 4 IN OCULAR LENS DEVELOPMENT AND FIBER CELL DIFFERENTIATION

Jin-Na Min, Yan Zhang, Demetrius Moskophidis, and Nahid F. Mivechi*

Institute of Molecular Medicine and Genetics, Molecular Chaperone Biology/Radiobiology Program, Medical College of Georgia, Augusta, GA, 30912.
*mivechi@immag.mcg.edu

INTRODUCTION Defects in the development and physiology of the eye lens as a result of gene mutations can cause cataracts, the commonest form of visual impairment in humans. Congenital cataracts account for around 10% of cases of childhood blindness, one-half of which have a genetic cause [1]. Ocular lens development is coordinated by expression of growth and transcription factors such as Pax6, FoxE3, Six3, Prox1, Sox2/3, Maf, Pitx3, AP-2a. Normally after formation of the lens vesicle, which is filled by elongated cells on its posterior surface (primary fibers), mitotically active cells from the monolayer of cuboidal epithelial cells at the anterior lens pole travel towards the equator where they elongate and differentiate into secondary lens fibers. Maturation of fiber cells is accompanied by

Processing biomedical texts

- **Named entity recognition**: difficult:
  - boundary detection:
    - capitalization patterns: often misleading
    - many premodifiers or postmodifiers that are part or not of the entity (**91 kDA protein**, activated **B cell** lines)
  - polysemous acronyms and terms: e.g., **PA** can stand for **pseudomonas aeruginosa**, **pathology** and **pulmonary artery**
  - synonymous acronyms and terms
- Supervised context dependent classification: HMM, CRF: often F_1 measure between 65-85%

[Zhang et al. BI 2004]

Processing biomedical texts

- **Entity relation recognition**:
  - Protein relation extraction
  - Literature based gene expression analysis
  - Determination of protein subcellular locations
  - Pathway prediction (cf. event detection)
    - methods relying on symbolic handcrafted rules, supervised (e.g., CRF) and unsupervised learning

[Stapley et al. PSBC 2002] [Glenisson et al. SIGKDD explorations 2003] [Friedman et al. BI 2001] [Huang et al. BI 2004] [Gaizauskas et al. ICNLP workshop 2000]

Заключение

- Обилие электронных документов, множество задач извлечения информации из текстов
- Есть прогресс в развитии методов извлечения
- Точность и полнота извлечения
  - зависят от набора шаблонов, размера обучающей коллекции
  - зависят друг от друга, поэтому для оценки используется F-мера

- Аккуратность (Accuracy), 2006
  В хорошо известных областях: Сущности – 90-98%,
  Атрибуты – 80%, Факты – 60-70%, События – 50-60%

- Дальнейшая проблема - сбор, хранение и обработка структурированной информации